SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 1-12, January 1989

OPTIMALITY CONDITIONS FOR SOME NONQUALIFIED PROBLEMS OF DISTRIBUTED CONTROL*

F. ABERGEL† AND R. TEMAM‡

Abstract. This article determines the necessary and sufficient optimality conditions for some nonqualified problems of optimal control, in the case of a distributed control for a system governed by a second-order elliptic partial differential equation. The authors study both bilateral and unilateral constraints on the state of the system.

Key words. distributed control, optimality conditions, nonqualified problems

AMS(MOS) subject classifications. 49B22, 49A22, 35K60

Introduction. One of the main difficulties in the study of optimal control problems is to give some necessary and sufficient conditions of optimality. For nonqualified problems, there is no standard way, even in the convex case, to give such conditions [L]. Nevertheless, the method developed by one of the authors for the study of variational problems in continuum mechanics [T] turns out to be very fruitful in the field of optimal control.

In this article we consider the following problems:

(**P**) Find (z, v) in $L^2(\Omega)\times L^2(\Omega)$ minimizing the cost function

$$J(z, v)=\left(\frac{1}{2\eta}\right)\int_\Omega v^2\,dx+\left(\frac{1}{2}\right)\int_\Omega |z-z_d|^2\,dx \tag{1.1.1}$$

with $(-\Delta z+z)=v$ in Ω, $z=0$ on $\partial\Omega$, $|z|\leqq\alpha$ almost everywhere in Ω.

The corresponding unilateral constraint problem reads:

(**R**) Find (z, v) in $L^2(\Omega)\times L^2(\Omega)$ minimizing the cost function

$$K(z, v)=\left(\frac{1}{2\eta}\right)\int_\Omega v^2\,dx+\left(\frac{1}{2}\right)\int_\Omega |z-z_d|^2\,dx \tag{2.1.1}$$

with $(-\Delta z+z)=v$ in Ω, $z=0$ on $\partial\Omega$, $z\leqq\alpha$ almost everywhere in Ω.

α and η are two strictly positive real numbers and z_d is given in $L^2(\Omega)$. In the case where Ω is a subset of the l-dimensional Euclidean space $\mathbb{R}^l$, $l=1, 2, 3$, Problems (**P**) and (**R**) correspond to the optimal heating of Ω: z is the temperature, v is the volumic heating (produced, for instance, by a laser beam), and z_d is the desired temperature. The constraint on z is a technological constraint, which can be easily interpreted as a no-burning condition.

In order to derive the system of optimality conditions for (**P**) and (**R**), we reformulate them as convex optimization problems in $H^2(\Omega)\cap H_0^1(\Omega)$, as follows. We let j, k be defined, respectively, as follows:

$$j(z)=\begin{cases}\left(\frac{1}{2}\right)\int_\Omega |z-z_d|^2\,dx+\left(\frac{1}{2\eta}\right)\int_\Omega |-\Delta z+z|^2\,dx & \text{if } z\in H^2(\Omega)\cap H_0^1(\Omega),\\ & |z|\leqq\alpha \text{ a.e. in } \Omega,\\ +\infty \quad \text{otherwise,}\end{cases}$$

* Received by the editors April 15, 1987; accepted for publication (in revised form) February 27, 1988.

† Laboratoire d'Analyse Numérique, Centre National de la Recherche Scientifique, Université Paris-Sud, 91405 Orsay, Cedex, France. Present address, Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802.

‡ Laboratoire d'Analyse Numérique, Centre National de la Recherche Scientifique, Université Paris-Sud, 91405 Orsay Cedex, France.

$$k(z)=\begin{cases}\left(\frac{1}{2}\right)\int_\Omega |z-z_d|^2\,dx+\left(\frac{1}{2\eta}\right)\int_\Omega |-\Delta z+z|^2\,dx & \text{if } z\in H^2(\Omega)\cap H_0^1(\Omega),\\ & \text{if } z\leqq\alpha \text{ a.e. in } \Omega,\\ +\infty \quad \text{otherwise.}\end{cases}$$

Therefore, (**P**) and (**R**) respectively, are equivalent to

$$(P)\qquad \text{(i)}\qquad \operatorname*{Inf}_{z\in H^2(\Omega)\cap H_0^1(\Omega)}\{j(z)\}$$

and

$$(R)\qquad \text{(ii)}\qquad \operatorname*{Inf}_{z\in H^2(\Omega)\cap H_0^1(\Omega)}\{k(z)\}.$$

If we wish to use the duality methods of convex analysis to derive the optimality conditions for (i), (ii), then we are led to the following dual problems of maximization:

$$(P^*)\qquad \text{(i')}\qquad \operatorname*{Sup}_{z\in H^2(\Omega)\cap H_0^1(\Omega)}\left\{\left(\frac{-\eta}{2}\right)\int_\Omega z^2\,dx+\left(\frac{1}{2}\right)\int_\Omega z_d^2\,dx-\int_\Omega \psi_\alpha(-\Delta z+z+z_d)\,dx\right\}$$

where

$$\psi_\alpha(s)=\frac{s^2}{s}\quad \text{if } |s|\leqq\alpha$$
$$=\alpha\left(|s|-\frac{\alpha}{2}\right)\quad \text{elsewhere,}$$

and

$$(R^*)\qquad \text{(ii')}\qquad \operatorname*{Sup}_{z\in H^2(\Omega)\cap H_0^1(\Omega)}\left\{\left(-\frac{\eta}{2}\right)\int_\Omega z^2\,dx+\left(\frac{1}{2}\right)\int_\Omega z_d^2\,dx-\int_\Omega \theta_\alpha(-\Delta z+z+z_d)\,dx\right\}$$

where

$$\theta_\alpha(s)=\frac{s^2}{s}\quad \text{if } s\leqq\alpha$$
$$=\alpha\left(s-\frac{\alpha}{2}\right)\quad \text{elsewhere.}$$

The crux of the matter is that neither (P^*) nor (R^*) is coercive on the natural space $X=H^2(\Omega)\cap H_0^1(\Omega)$. Thus we must investigate the existence of solutions to (i') and (ii') in a larger space than X, in order to recover coercivity.

For (i'), the natural space is $\mathrm{BL}_0(\Omega)=\{u\in L^2(\Omega),\ (-\Delta u+u)$ is a bounded measure on Ω, $u=0$ on $\partial\Omega\}$ due to the linear behaviour of ψ_α at infinity. We prove that (P^*) has (generalized) solutions in $\mathrm{BL}_0(\Omega)$, and extend the classical optimality conditions to them.

As for (ii'), the problem is slightly more delicate, for (R^*) cannot be extended to the whole space $\mathrm{BL}_0(\Omega)$, due to the quadratic part in $\mathcal{C}_\alpha$. However, we prove that such an extension is possible if we restrict ourselves to the functions u in $\mathrm{BL}_0(\Omega)$ such that $(-\Delta u+u)$ is in a convex cone of the space of bounded measures on Ω. We then establish the system of optimality conditions.

We use the following classical notation: $W^{m,p}(\Omega)$ for the Sobolev space of order m on $L^p(\Omega)$, $W_0^{m,p}$ for the closure in $W^{m,p}(\Omega)$ of the Schwartz class $\mathcal{D}(\Omega)$, and $\mathcal{D}'(\Omega)$ for the dual space of the latter; we also classically write $H^m(\Omega)$ for $W^{m,2}(\Omega)$, and $M_1(\Omega)$ for the space of bounded measures on Ω (see [Ad], [LM], [B]).

1. A distributed control problem with bilateral constraints.

1.1. Variational formulation. Let Ω be an open bounded set of $\mathbb{R}^N$, whose boundary Γ is a compact C^∞, $(N-1)$-dimensional manifold, Ω being locally on one side of Γ. We want to study the following problem of optimal control.

(**P**) Find (z, v) in $L^2(\Omega)\times L^2(\Omega)$ minimizing the cost function

$$J(z, v)=\left(\frac{1}{2\eta}\right)\int_\Omega v^2\,dx+\left(\frac{1}{2}\right)\int_\Omega |z-z_d|^2\,dx \tag{1.1.1}$$

z being such that $(-\Delta z+z)=v$ in Ω, $z=0$ on Γ, $|z|\leqq\alpha$ a.e. on Ω.

z_d is given in $L^2(\Omega)$, and α, η are two strictly positive real numbers. The conditions on z imply that it belongs to $H^2(\Omega)\cap H_0^1(\Omega)$ [LM] and the variational formulation of (P) is then

$$(P)\qquad \operatorname*{Inf}_{z\in H^2(\Omega)\cap H_0^1(\Omega)}\{j(z)\}, \tag{1.1.2}$$

the functional j being defined by

$$j(z)=\begin{cases}\dfrac{1}{2}\displaystyle\int_\Omega |z-z_d|^2\,dx+\frac{1}{2}\eta\int_\Omega |-\Delta z+z|^2\,dx\\ \qquad\qquad \text{if } z\in X=H^2(\Omega)\cap H_0^1(\Omega),\quad |z|\leqq\alpha \text{ a.e. on } \Omega,\\ +\infty \quad \text{otherwise.}\end{cases} \tag{1.1.3}$$

We have Proposition 1.1.1 below (see [ET]).

PROPOSITION 1.1.1. *There exists a unique optimal state z for Problem (P), which is the only solution of (P).*

We are now going to give the expression of the dual problem (P^*) of (P), in order to study the system of optimality conditions for Problem (**P**).

1.2. Duality. We shall use the duality methods described in [ET].

Let Y be the space $(L^2(\Omega))^2$; we define the operator $\wedge$, from X into Y, by

$$\wedge z=(z, -\Delta z+z). \tag{1.2.1}$$

Problem (P) has the following form:

$$(P)\qquad \operatorname*{Inf}_{z\in X}\{F(z)+G(\wedge z)\} \tag{1.2.2}$$

where F, G are defined as follows: $F\equiv 0$; $G(p)=G_1(p_1)+G_2(p_2)$, with

$$G_1(p_1)=\begin{cases}\dfrac{1}{2}\displaystyle\int_\Omega |p_1-z_d|^2\,dx & \text{if } |p_1|\leqq\alpha \text{ a.e. on } \Omega,\\ +\infty \quad \text{otherwise,}\end{cases} \tag{1.2.3}$$

$$G_2(p_2)=\frac{1}{2\eta}\int_\Omega |p_2|^2\,dx. \tag{1.2.4}$$

The dual problem (P^*) of (P) is then

$$(P^*)\qquad \operatorname*{Sup}_{q\in Y}\{-F^*(\wedge^* q)-G^*(-1)\}, \tag{1.2.5}$$

F^* (respectively, G^*) being the convex conjugate function of F (respectively, G), and $\wedge^*$, the transposed operator of $\wedge$.

Remark 1.2.1. Y is identified with its conjugate Y^*, thanks to its Hilbertian structure.

Let us now compute the expression of F^* and G^*. For G^*, we easily find

$$G^*(p) = G_1^*(p_1) + G_2^*(p_2) \tag{1.2.6}$$

where

$$G_1^*(p_1) = -\frac{1}{2}\int_\Omega z_d^2\, dx + \int_\Omega \Psi_\alpha(p_1 + z_d)\, dx, \tag{1.2.7}$$

$$G_2^*(p_2) = \frac{\eta}{2}\int_\Omega |p_2|^2\, dx, \tag{1.2.8}$$

and the function Ψ_α is defined by

$$\Psi_\alpha(s) = \begin{cases} s^2/2 & \text{if } |s| \leqq \alpha, \\ \alpha(|s| - \alpha/2) & \text{if } |s| \geqq \alpha. \end{cases} \tag{1.2.9}$$

For F^*, we write

$$\begin{aligned} F^*(\wedge^* p) &= \operatorname*{Sup}_{z \in X} \langle p, \wedge z \rangle \\ &= \operatorname*{Sup}_{z \in X} \left(\int_\Omega [p_1 \cdot z + (-\Delta z + z) \cdot p_2]\, dx \right) \\ &\geqq \operatorname*{Sup}_{z \in \mathcal{D}(\Omega)} \left(\int_\Omega [p_1 \cdot z + (-\Delta z + z) \cdot p_2]\, dx \right) \\ &\geqq \begin{cases} 0 & \text{if } -\Delta p_2 + p_2 = -p_1 \text{ in } \Omega, \\ +\infty & \text{otherwise.} \end{cases} \end{aligned} \tag{1.2.10}$$

If p_2 is such that $(-\Delta p_2 + p_2)$ belongs to $L^2(\Omega)$, we can define its boundary values $(p, \partial p/\partial \nu)$ as distributions on Γ (see [LM]). Moreover, the following Green formula holds for z in $H^2(\Omega)$:

$$\int_\Omega (-\Delta z + z) \cdot p_2\, dx = \int_\Omega (-\Delta p_2 + p_2) \cdot z\, dx - \int_\Gamma \left(\left(\frac{\partial z}{\partial \nu} \right) p_2 - \left(\frac{\partial p_2}{\partial \nu} \right) \cdot z \right) d\Gamma \tag{1.2.11}$$

(the integral on Γ being, in fact, a duality product in the suitable distributions spaces on Γ).

From (1.2.11), we easily deduce the expression of F^*:

$$F^*(\wedge^* p) = \begin{cases} 0 & \text{if } -\Delta p_2 + p_2 = 0 \text{ in } \Omega, \quad p_2 = 0 \text{ on } \Gamma, \\ +\infty & \text{otherwise.} \end{cases} \tag{1.2.12}$$

After eliminating p_1, we find the following for (P^*):

$$(P^*) \qquad \operatorname*{Sup}_{p \in X} \left\{ \left(\frac{-\eta}{2} \right) \int_\Omega p^2\, dx + \left(\frac{1}{2} \right) \int_\Omega z_d^2\, dx - \int_\Omega \Psi_\alpha(-\Delta p + p + z_d)\, dx \right\}. \tag{1.2.13}$$

By lack of coercivity (see the definition of Ψ_α), we do not know whether Problem (P^*) has a solution in X. In the following sections, we shall show how it is possible to overcome this difficulty, by extending the class of admissible elements for Problem (P^*). We shall also give the necessary *and* sufficient conditions for an admissible state z, for problem P, to be the optimal state z.

For the moment, we give the following comparison result.

PROPOSITION 1.2.1. *The extrema of* (P) *and* (P^*) *are equal*:

$$-\infty < \operatorname{Sup}(P^*) = \operatorname{Inf}(P).$$

The proof is omitted. We show that Problem (P) is normal [ET].

1.3. The generalized problem. Definition (1.2.9) of Ψ_α shows that it is natural to look for solutions of (P^*) in the space $\mathrm{BL}_0(\Omega)$ defined by

$$\mathrm{BL}_0(\Omega) = \{u \in L^2(\Omega), (-\Delta u + u) \in M_1(\Omega), u = 0 \text{ on } \Gamma\} \tag{1.3.1}$$

where $M_1(\Omega)$ is the space of bounded measures on Ω (BL stands for "Bounded Laplacian"; the index 0 refers to the Dirichlet conditions on Γ). $\mathrm{BL}_0(\Omega)$ is a Banach space for the norm

$$\|u\|_{\mathrm{BL}_0(\Omega)} \equiv \|u\|_{L^2(\Omega)} + \|-\Delta u + u\|_{M_1(\Omega)}. \tag{1.3.2}$$

Moreover (see [M]), the space $\mathrm{BL}_0(\Omega)$ is continuously imbedded in the space

$$W_0^{1,s}(\Omega) \quad \text{for } 1 \leqq s < \frac{N}{N-1}.$$

In order to extend Problem (P^*) to $\mathrm{BL}_0(\Omega)$, we need to define $\Psi_\alpha(\mu)$, when μ is a bounded measure, and Ψ_α is defined by (1.2.9). We shall refer extensively to the results of [DT1], [DT2], [T], and recall what will be useful to us in Proposition 1.3.1.

PROPOSITION 1.3.1. *Let* Ψ *be a convex function of one real variable, let* Ψ^* *be its conjugate function, and let* Ψ_∞ *be its asymptotic function. We suppose that*

(i) *There exists* C, *such that for all* $s \in \mathbb{R}$, $|\Psi(s)| \leqq C(1+|s|)$;

(ii) Ψ^* *is bounded on its domain* K;

(iii) $\Psi \geqq 0, \Psi(0) = 0$.

Let μ *be a bounded measure having the decomposition* $\mu = h \cdot dx + \theta_s \cdot |\mu_s|$ *with respect to the Lebesgue measure* dx, μ_s *singular with respect to* dx [B]. *The measure* $\Psi(\mu)$ *is then defined as follows*:

$$\Psi(\mu) = (\Psi \circ h) \cdot dx + \Psi_\infty(\theta_s) \cdot |\mu_s|. \tag{1.3.3}$$

Furthermore, we have the duality formula

$$\forall \Phi \in \mathscr{C}_0(\Omega) \quad \langle \Psi(\mu), \Phi \rangle = \operatorname{Sup}\left\{ \int_\Omega \Phi \cdot g \, d\mu - \int_\Omega \Phi \cdot \Psi^*(g) \, dx \right\}, \tag{1.3.4}$$

the supremum being taken for $g \in \mathscr{C}_0(\Omega)$, $\Psi^*(g) \in L^1(\Omega)$; *and relation* (1.3.4) *is still valid for* $\Phi \in \mathscr{C}(\Omega)$, $\Phi \geqq 0$.

It is now obvious how to formulate the generalized problem (Q^*):

$$(Q^*) \qquad \operatorname*{Sup}_{u \in \mathrm{BL}_0(\Omega)} \left\{ \left(\frac{-\eta}{2}\right) \int_\Omega u^2 \, dx + \left(\frac{1}{2}\right) \int_\Omega z_d^2 \, dx - \int_\Omega \Psi_\alpha(-\Delta u + u + z_d) \right\}. \tag{1.3.5}$$

In the expression above, $\int_\Omega \Psi_\alpha(-\Delta u + u + z_d)$ represents the total mass of the bounded measure $\Psi_\alpha(-\Delta u + u + z_d)$. We obviously have the following inequality:

$$\operatorname{Sup}(P^*) \leqq \operatorname{Sup}(Q^*). \tag{1.3.6}$$

In § 4 we shall introduce a generalized duality, between (P) and (Q^*), that will enable us to prove that (1.3.6) is, in fact, an equality, and will also give the system of optimality conditions for Problem (P).

1.4. Generalized duality. Our purpose in this section is to give meaning to the expression "$(-\Delta u + u) \cdot z$" when z is in X, u is in $\mathrm{BL}_0(\Omega)$, and z is admissible for (P).

Actually, we are going to prove that $(-\Delta u+u)\cdot z$ can be defined as a bounded measure on Ω, and that there exists a generalized Green formula for z and u. More precisely, we have the following result.

PROPOSITION 1.4.1. *Let (z, u) belong to $X\times \mathrm{BL}_0(\Omega)$, z being admissible for Problem (P). We can define a distribution, denoted $(-\Delta u+u)\cdot z$, by the following formula:*

$$\forall \Phi\in \mathscr{D}(\Omega)\quad \langle(-\Delta u+u)\cdot z,\Phi\rangle=\int_\Omega [-\Delta(z\cdot\Phi)+z\cdot\Phi]\cdot u\,dx. \tag{1.4.1}$$

The distribution defined by (1.4.1) satisfies the following:

(i) *$(-\Delta u+u)\cdot z$ is a bounded measure on Ω.*

(ii) *We have the equality*

$$\int_\Omega (-\Delta u+u)\cdot z=\int_\Omega (-\Delta z+z)\cdot u\,dx. \tag{1.4.2}$$

Proof. We first remark that the right-hand side of (1.4.1) makes sense for u in $L^2(\Omega)$ and z in $X=H^2(\Omega)\cap H_0^1(\Omega)$. Let us now choose a sequence z_n of smooth functions approximating z in the following sense:

$$z_n \text{ tends to } z \quad \text{in } X, \tag{1.4.3}$$

$$\|z_n\|_{L^\infty(\Omega)}\leqq \|z\|_{L^\infty(\Omega)}. \tag{1.4.4}$$

Such a sequence is classically obtained, for instance, by solving the following problem associated to the heat equation:

$$\begin{aligned} &\partial v/\partial t-\Delta v=0 && \text{in } \Omega\times]0,\infty[,\\ &v=0 && \text{on } \Gamma\times]0,\infty[,\\ &v(x,0)=z(x) && \text{in } \Omega, \end{aligned} \tag{1.4.5}$$

and setting $z_n=v(\cdot, t_n)$, t_n being a sequence of strictly positive real numbers converging to zero. Frequently (1.4.4) is then a consequence of the maximum principle for second-order parabolic equations.

We now set $T_n=(-\Delta u+u)\cdot z_n$; T_n is a bounded measure satisfying

$$\|T_n\|_{M_1(\Omega)}\leqq \|z\|_{L^\infty(\Omega)}\cdot\|-\Delta u+u\|_{M_1(\Omega)}. \tag{1.4.6}$$

Moreover, T_n obviously converges to $(-\Delta u+u)\cdot z$ in $\mathscr{D}'(\Omega)$, and this proves assertion (i). We also remark that T_n converges to $(-\Delta u+u)\cdot z$ in the sense of the tight convergence of measures. That is true because the sequence (T_n) satisfies the Prokhorov condition [B]:

$$\forall\varepsilon>0\quad \exists K_\varepsilon\quad \forall n\quad \int_{\Omega\setminus K_\varepsilon}|T_n|<\varepsilon \tag{1.4.7}$$

where K_ε is a Borel subset of Ω. As a matter of fact, $(-\Delta u+u)$ is a bounded measure, and the functions z_n are uniformly bounded on Ω. Therefore, the sequence T_n is relatively compact for the topology of the tight convergence in $M_1(\Omega)$.

To prove (ii), let us now choose a function Φ in $\mathscr{C}^\infty(\bar\Omega)$; thanks to the results of [LM], we have the Green formula

$$\begin{aligned} \int_\Omega (-\Delta u+u)\cdot z_n\cdot\Phi=\int_\Omega [-\Delta(z_n\cdot\Phi)+z_n\cdot\Phi]\cdot u\,dx-\langle\partial u/\partial\nu, z_n\cdot\Phi\rangle_{\mathscr{D}(\Gamma)\times\mathscr{D}'(\Gamma)}\\ +\langle u,\partial(z_n\Phi)/\partial\nu\rangle_{\mathscr{D}(\Gamma)\times\mathscr{D}'(\Gamma)}. \end{aligned} \tag{1.4.8}$$

Taking into account the conditions ($u=0$ on Γ) and ($z_n=0$ on Γ), we obtain

$$(1.4.9)\qquad \forall n\in\mathbb{N},\quad \forall \Phi\in\mathscr{C}^\infty(\bar{\Omega})\quad \int_\Omega (-\Delta u+u)\cdot z_n\cdot\Phi=\int_\Omega [-\Delta(z_n\cdot\Phi)+z_n\cdot\Phi]\cdot u\,dx.$$

Assertion (ii) is then proved by letting n tend to ∞ in (1.4.9), with $\Phi\equiv 1$, and using the tight convergence of T_n to $(-\Delta u+u)\cdot z$. □

1.5. Existence of solution of (Q^*) in $BL_0(\Omega)$; system of optimality conditions for (P). In this section, we give our final results for the study of Problem (P), namely, the existence of an adjoint state in $BL_0(\Omega)$ for the optimal state z, and the system of optimality conditions related to that problem.

We start with a lemma.

LEMMA 1.5.1. *Let J^* be the functional defined on $BL_0(\Omega)$ by*

$$(1.5.1)\qquad J^*(u)=\frac{\eta}{2}\int_\Omega u^2\,dx-\frac{1}{2}\int_\Omega z_d^2\,dx+\int_\Omega \Psi_\alpha(-\Delta u+u+z_d).$$

Then we have the following:

(i) *J^* is lower semicontinuous on $BL_0(\Omega)$ for the weak topology τ_1 ($u_n\to u$ for τ_1 if $u_n\to u$ in $L^2(\Omega)$ weakly and $(-\Delta u_n+u_n)\to(-\Delta u+u)$ in $M_1(\Omega)$ vaguely).*

Moreover, any bounded set of $BL_0(\Omega)$ is relatively compact for the τ_1 topology.

(ii) *If (z,u) belongs to $X\times BL_0(\Omega)$, z being admissible for Problem (P), we have*

$$(1.5.2)\qquad (-J^*(u))\leqq j(z)-\left(\frac{\eta}{2}\right)\int_\Omega\left|\left(\frac{1}{\eta}\right)(-\Delta z+z)+u\right|^2 dx.$$

Proof. We first notice that (i) is a consequence of the definition of $BL_0(\Omega)$ and the properties of $\Psi(\mu)$[DT1].

For (ii), we use the sequence z_n above and the duality formula (1.3.4) to derive the following inequalities:

$$\begin{aligned}\int_\Omega \Psi_\alpha(-\Delta u+u+z_d)&\geqq\int_\Omega(-\Delta u+u+z_d)\cdot z_n-\int_\Omega\Psi_\alpha^*(z_n)\,dx\\&\geqq\left(\text{for }\Psi_\alpha^*(s)=\frac{1}{2}s^2,\ \forall s\in\mathbb{R},\ |s|\leqq\alpha\right)\\&\geqq\int_\Omega(-\Delta u+u+z_d)\cdot z_n-\frac{1}{2}\int_\Omega z_n^2\,dx\\&\geqq(\text{for } T_n \text{ converges tightly to } (-\Delta u+u)\cdot z)\\&\geqq\int_\Omega(-\Delta u+u+z_d)\cdot z-\frac{1}{2}\int_\Omega z^2\,dx.\end{aligned}$$

We now use the Green formula (1.4.2) to obtain

$$\begin{aligned}J^*(u)&\geqq\frac{\eta}{2}\int_\Omega u^2\,dx-\frac{1}{2}\int_\Omega z_d^2\,dx-\frac{1}{2}\int_\Omega z^2\,dx+\int_\Omega(-\Delta z+z)\cdot u\,dx+\int_\Omega z_d\cdot z\,dx\\&\geqq-j(z)+\frac{\eta}{2}\int_\Omega\left|\frac{1}{\eta}(-\Delta z+z)+u\right|^2 dx,\end{aligned}$$

and Lemma 1.5.1 is proved. □

Its fundamental importance is due to the fact that it makes possible the use of the standard method of calculus of variations to study Problem (Q^*) in $BL_0(\Omega)$. Our results are summed up in Theorem 1.5.2.

THEOREM 1.5.2.

(i) *We have the equality* $\operatorname{Inf}(P) = \operatorname{Sup}(Q^*)$.

(ii) *There exists, in* $BL_0(\Omega)$, *an adjoint state for the optimal state* z: *it is a solution of* (Q^*).

(iii) *The necessary and sufficient conditions for a couple* (z, u) *of admissible elements for* (P) *and* (Q^*) *to be an optimal couple are*:

$$\int_\Omega \Psi_\alpha(-\Delta u + u + z_d) = \int_\Omega (-\Delta u + u + z_d)\cdot z - \frac{1}{2}\int_\Omega z^2\,dx, \tag{1.5.3}$$

$$u = -\frac{1}{\eta}(-\Delta z + z) \quad \text{a.e. in } \Omega. \tag{1.5.4}$$

Proof. Theorem 1.5.2 follows directly from Lemma 1.5.1. □

2. A distributed control problem with a unilateral constraint. The purpose in § 2 is to extend our results to the unilateral case, where the constraint on the state z has the form ($z \leqq \alpha$ almost everywhere on Ω). The main differences with the bilateral problem **P** come from the fact that an admissible state z no longer belongs to $L^\infty(\Omega)$, and that the analogue of the function Ψ_α (1.2.9) does not satisfy condition (i) of Proposition 1.3.1, i.e., that it be at most linear at infinity. Nevertheless, we shall see that it is possible to overcome these new difficulties, with appropriate methods, and to prove a result as complete as Theorem 1.5.2.

2.1. Variational formulation. Primal and dual problems. The geometrical assumptions being as in § 1, we turn to the study of the following problem.

(**R**) Find (z, v) in $L^2(\Omega)\times L^2(\Omega)$ minimizing the cost function

$$K(z, v) = \frac{1}{2\eta}\int_\Omega v^2\,dx + \frac{1}{2}\int_\Omega |z - z_d|^2\,dx \tag{2.1.1}$$

z being such that $(-\Delta z + z) = v$ in Ω, $z = 0$ on Γ, $z \leqq \alpha$ almost everywhere on Ω.

z_d is given in $L^2(\Omega)$, and α, δ are two strictly positive real numbers. The variational formulation of (R) is

$$(R) \qquad \operatorname*{Inf}_{z\in X}\{k(z)\} \tag{2.1.2}$$

where the functional k is defined by

$$k(z) = \begin{cases} \left(\frac{1}{2}\right)\int_\Omega |z - z_d|^2\,dx + \frac{1}{2\eta}\int_\Omega |-\Delta z + z|^2\,dx \\ \qquad\qquad\qquad\qquad \text{if } z \in X, \quad z \leqq \alpha \quad \text{a.e. on } \Omega, \\ +\infty \quad \text{otherwise.} \end{cases} \tag{2.1.3}$$

As in the bilateral case, we easily see the following [ET].

PROPOSITION 2.1.1. *There exists a unique optimal state* z' *for* R; *it is the minimizer of* (R).

The dual problem (R^*) of (R) is obtained by the same methods as before, and its formulation is

$$(2.1.4) \qquad (R^*) \qquad \operatorname*{Sup}_{p\in X}\left\{\left(\frac{-\eta}{2}\right)\int_\Omega p^2\,dx+\frac{1}{2}\int_\Omega z_d^2\,dx-\int_\Omega \Theta_\alpha(-\Delta p+p+z_d)\,dx\right\}$$

where the function Θ_α is defined by

$$(2.1.5) \qquad \Theta_\alpha(s)=\begin{cases} s^2/2 & \text{if } s\leqq\alpha,\\ \alpha(s-\alpha/2) & \text{if } s\geqq\alpha.\end{cases}$$

We can show that the extrema of (R) and (R^*) are equal, but we do not know whether Problem (R^*) has a solution in X.

We want to extend Problem (R^*) to $\mathrm{BL}_0(\Omega)$, and the new difficulty is that we cannot define $\Theta_\alpha(\mu)$ for any bounded measure μ, due to the quadratic behaviour of Θ_α at infinity. We recall in the following proposition the results of [DT2] that are necessary to extend Problem (R^*) to $\mathrm{BL}_0(\Omega)$.

PROPOSITION 2.1.2. *Let Θ_α be defined as in* (2.1.5), *and μ be a bounded measure on Ω. We suppose that μ admits the Lebesgue decomposition $\mu=h\cdot dx+\theta_s\cdot|\mu_s|$, μ_s singular with respect to the Lebesgue measure and such that*

(i) *μ_s is positive (i.e., $\theta_s\equiv 1$ in Ω).*

(ii) *$h^-=-\mathrm{Inf}(h,0)$ is in $L^2(\Omega)$.*

Then the bounded measure $\Theta_\alpha(\mu)$ is defined by

$$(2.1.6) \qquad \Theta_\alpha(\mu)(\Theta_\alpha\circ h)\cdot dx+\Theta_{\alpha,\infty}(\theta_s)\cdot|\mu_s|$$

with $\Theta_{\alpha,\infty}(s)=0$ if $s<0$, and $+\infty$ if $s\geqq 0$, and we have the duality formula

$$(2.1.7) \qquad \forall\Phi\in\mathscr{C}_0(\Omega)\quad \langle\Theta_\alpha(\mu),\Phi\rangle=\operatorname{Sup}\left\{\int_\Omega \Phi\cdot g\cdot d\mu-\int_\Omega \Phi\cdot\Theta_\alpha^*(g)\cdot dx\right\},$$

the supremum being taken for $g\in\mathscr{C}_0(\Omega)$, $g\leqq\alpha$ in Ω. Moreover, relation (2.1.7) *is still valid for $\Phi\in\mathscr{C}(\Omega)$, $\Phi\geqq 0$ in Ω.*

(These results come from Theorem 2.1 of [DT2].)

Remark 2.1.1. In this particular case, the expression of $\Theta_\alpha(\mu)$ is

$$(2.1.8) \qquad \Theta_\alpha(\mu)=(\Theta_\alpha\circ h)\cdot dx+\alpha\cdot\mu_s.$$

We define the set $M^\alpha(\Omega)=\{\mu\in M_1(\Omega),\ \mu$ satisfies (i) and (ii)$\}$. One of its main properties [DT2, Lemma 3.2.1] is the following.

(A) If μ_n is a sequence in $M^\alpha(\Omega)$ such that $\Theta_\alpha(\mu_n)$ is bounded, and if μ_n converges vaguely to a measure μ, then μ belongs to $M^\alpha(\Omega)$. Furthermore, we have

$$\int_\Omega \Theta_\alpha(\mu)\leqq\varliminf_{n\to\infty}\int_\Omega \Theta_\alpha(\mu_n).$$

We can now give the expression of the generalized problem (S^*):

$$(2.1.9) \qquad (S^*) \qquad \operatorname{Sup}\left\{\left(\frac{-\eta}{2}\right)\int_\Omega u^2\,dx+\frac{1}{2}\int_\Omega z_d^2\,dx-\int_\Omega \Theta_\alpha(-\Delta u+u+z_d)\right\},$$

where the supremum is taken for $u\in\mathrm{BL}_0(\Omega)$, $(-\Delta u+u+z_d)\in M^\alpha(\Omega)$.

Obviously we have

$$(2.1.10) \qquad \mathrm{Inf}(R)\leqq\mathrm{Sup}(S^*).$$

Our purpose in the next section will be to define a generalized duality between (R) and (S^*), so as to give the system of optimality conditions for Problem (**R**).

2.2. Green's formula and duality. If z and u are admissible, respectively, for (R) and (S^*), we can see that it is no longer possible to define $(-\Delta u+u)\cdot z$ as a bounded measure by using the method of Proposition 1.4.1, for z is not bounded in Ω. However, we can prove a very similar result.

PROPOSITION 2.2.1. *Let (z, u) be a couple of admissible elements, respectively, for (R) and (S^*). The distribution $(-\Delta u+u)\cdot z$ is a bounded measure in Ω, and the Green formula* (1.4.2) *still holds.*

Proof. Let us admit for the moment the following result.

LEMMA 2.2.2. *Let (y, v) belong to $X\times \mathrm{BL}_0(\Omega)$, such that $(-\Delta v+v)$ is a positive measure; then the distribution $(-\Delta v+v)\cdot y$ defined in* (1.4.1) *is a bounded measure, and we have the equality*

$$\int_\Omega (-\Delta v+v)\cdot y=\int_\Omega (-\Delta y+y)\cdot v\,dx. \tag{2.2.1}$$

We can now prove Proposition 2.2.1. We write $u=u_1-u_2$, where u_1 and u_2 are obtained by solving the Dirichlet problems corresponding, respectively, to $(-\Delta u+u+z_d)^+$ and $((-\Delta u+u+z_d)^-+z_d)$. Then $(-\Delta u_1+u_1)\cdot z$ is defined as a measure by using Lemma 2.2.2, $(-\Delta u_2+u_2)\cdot z$ has a natural meaning (thanks to the assumptions on u), and the Green formula comes from (2.2.1) and the classical formula in X. That proves Proposition 2.2.1. □

Proof of Lemma 2.2.2. We start with the case where y has a constant sign, say $y\leqq 0$, and use a sequence y_n of smooth functions satisfying the following conditions:

(i) y_n converges to y in X;

(ii) For all $n\in\mathbb{N}$, $0\geqq y_n\geqq \operatorname{Inf} y$ in Ω.

Such a sequence is obtained in exactly the same way as the sequence z_n of Proposition 1.4.1. We set $T'_n=(-\Delta v+v)\cdot y_n$; T'_n is a bounded negative measure (for y_n is smooth in $\bar\Omega$), and we have the equality

$$\int_\Omega (-\Delta v+v)\cdot y_n=\int_\Omega (-\Delta y_n+y_n)\cdot v\,dx \tag{2.2.2}$$

(see the proof of Proposition 1.4.1). The right-hand side of (2.2.2) has a limit because of (i); moreover, due to the assumptions on the signs of $(-\Delta v+v)$ and y_n, the left-hand side of (2.2.1) is the opposite of the norm of $(-\Delta v+v)\cdot y_n$ in $M_1(\Omega)$. Hence, we can ensure that the sequence T'_n is bounded in $M_1(\Omega)$. Therefore, $(-\Delta v+v)\cdot y$ is a bounded measure on Ω, as the limit in $\mathscr{D}'(\Omega)$ of a bounded sequence in $M_1(\Omega)$.

Now one thing is left to prove, namely, the tight convergence of T'_n to $(-\Delta v+v)\cdot y$. As a matter of fact, that will be sufficient to prove (2.2.1). Let us consider the linear functional T''_n, defined on $\mathscr{C}(\bar\Omega)$ by

$$\langle T''_n,\Phi\rangle=\int_\Omega [-\Delta v+v]y_n\cdot\Phi. \tag{2.2.3}$$

For the same reasons as in Proposition 1.4.1, we have, for Φ in $\mathscr{C}^\infty(\bar\Omega)$:

$$\langle T''_n,\Phi\rangle=\int_\Omega [-\Delta(y_n\cdot\Phi)+y_n\cdot\Phi]\cdot v\,dx. \tag{2.2.4}$$

Hence, $(T''_n)_{n\in\mathbb{N}}$ is a sequence of negative linear functionals on $\mathscr{C}(\bar\Omega)$, which is bounded in $[\mathscr{C}(\bar\Omega)]'$ (use (2.2.4) with $\Phi\equiv 1$). Thus, there exists a subsequence, still denoted by T''_n, which is vaguely convergent to a functional T'' in $[\mathscr{C}(\bar\Omega)]'$. As $\bar\Omega$ is compact, the vague and tight convergences are equivalent, for positive measures on $\bar\Omega$. We now use this result from [B]:

Result. The canonical injection **i** from the positive cone of $M_1(\Omega)$ onto its image in $[\mathscr{C}(\bar{\Omega})]'$ is a homeomorphism when each space is endowed with the topology of tight convergence.

As we obviously have $\mathbf{i}(T'_n) = T''_n$, the tight convergence of T'_n to $[-\Delta v + v] \cdot y$ is proved.

We now return to the general case: when y is in X, we can always write it as the difference of two positive elements of X. For instance, we can solve the Dirichlet problems relative to the positive and negative parts of $(-\Delta y + y)$. Thanks to the maximum principle, the corresponding functions y_1 and y_2 are positive, and we obviously have $y = y_1 - y_2$; we then set $[-\Delta v + v] \cdot y = [-\Delta v + v] \cdot y_1 - [-\Delta v + v] \cdot y_2$ and, using the results above, define $[-\Delta v + v] \cdot y$ as a bounded measure. This definition of $[-\Delta v + v] \cdot y$ is independent of the decomposition of y as the difference of two positive functions in X; that is a consequence of (2.2.4), and of the density of $\mathscr{C}^\infty(\Omega)$ in $\mathscr{C}(\Omega)$. Hence Lemma 2.2.2 is proved. □

Remark 2.2.2. The application $y \to [-\Delta v + v] \cdot y$, for fixed v, is continuous from X endowed with its strong topology, into $M_1(\Omega)$ endowed with the topology of tight convergence. That is also a consequence of (2.2.4), and of the density of $\mathscr{C}^\infty(\Omega)$ in $\mathscr{C}(\Omega)$.

We now have all the technical elements required to solve Problem (R) completely. Before giving our final results, we state a lemma.

LEMMA 2.2.3. *Let K^* be the functional defined in* $\mathrm{BL}_0(\Omega)$ *by*

$$K^*(u) = \begin{cases} \left(\frac{\eta}{2}\right) \int_\Omega u^2 \, dx - \frac{1}{2} \int_\Omega z_d^2 \, dx + \int_\Omega \Theta_\alpha(-\Delta u + u + z_d) \\ \qquad \text{if } u \in \mathrm{BL}_0(\Omega),\ (-\Delta u + u + z_d) \in M^\alpha(\Omega), \\ +\infty \quad \text{otherwise.} \end{cases} \tag{2.2.5}$$

K^* enjoys the following properties:

(i) *If (z, u) is a couple of admissible elements respectively, for (R) and (S^*), we have*

$$(-K^*(u)) \leqq k(z) - \frac{\eta}{2} \int_\Omega \left| \frac{1}{\eta} (-\Delta z + z) + u \right|^2 dx. \tag{2.2.6}$$

(ii) *Any sequence u_n in* $\mathrm{BL}_0(\Omega)$ *such that $K^*(u_n)$ is bounded has a cluster point u for the τ_1 topology of* $\mathrm{BL}_0(\Omega)$, *and we have*

$$K^*(u) \leqq \varliminf_{n \to \infty} (K^*(u_n)). \tag{2.2.7}$$

Proof. The proof is omitted. It is definitely similar to that of Lemma 1.5.1 because of the duality formula (2.1.7) and Proposition 2.2.1.

We now conclude the study of the optimality system for Problem (R).

THEOREM 2.2.4. (i) *The extrema of (R) and (S^*) are equal.*

(ii) *There exists a solution of Problem (S^*) in* $\mathrm{BL}_0(\Omega)$.

(iii) *The necessary and sufficient conditions for a couple (z, u) of admissible elements to be an optimal couple are*

$$\int_\Omega \Theta_\alpha(-\Delta u + u + z_d) = \int_\Omega (-\Delta u + u + z_d) \cdot z - \frac{1}{2} \int_\Omega z_d^2 \, dx, \tag{2.2.8}$$

$$u = \left(\frac{1}{\eta}\right)(-\Delta z + z) \quad \text{a.e. in } \Omega. \tag{2.2.9}$$

Proof. The proof is the same as that of Theorem 1.5.2.

REFERENCES

[A] F. ABERGEL, *A non well-posed problem in convex optimal control*, Thèse de doctorat, Université d'Orsay, Orsay, France, 1986.

[AD] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[B] N. BOURBAKI, *Integration*, Hermann, Paris, 1963, Chap. 9.

[DT1] F. DEMENGEL AND R. TEMAM, *Convex function of a measure and applications*, Indiana Math. J., 33 (1984), pp. 673-709.

[DT2] ———, *Convex function of a measure and applications: the unbounded case*, in FERMAT *Days* 85; *Mathematics for Optimization*, J. B. Hiriart-Urruty, ed., Elsevier-North-Holland, 1986, pp. 103-134.

[ET] I. EKELAND AND R. TEMAM, *Analyse convexe et problemes variationnels*, Dunod Gauthier-Villars, Paris, 1973.

[L] J-L. LIONS, *Some Methods in the Mathematical Analysis of Systems and their Control*, Gordon and Breach, New York, 1981.

[LM] J-L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Dunod, Paris, 1968.

[M] C. MIRANDA, *Partial Differential Equations of Elliptic Type*, Second edition, Springer-Verlag, Berlin, New York, 1970.

[T] R. TEMAM, *Mathematical Problems in Plasticity*, Gauthier-Villars, Paris, New York, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 13-26, January 1989

002

CONVERGENCE OF SQP-LIKE METHODS FOR CONSTRAINED OPTIMIZATION*

STEPHEN WRIGHT†

Abstract. The problem of constrained optimization

$$\min f(x) \quad \text{s.t. } x \in \Omega$$

is sometimes solved by an iterative method, in which Ω is replaced by some other set $\Omega(x_k)$ with simple geometry at each iteration x_k. Sequential quadratic programming methods for nonlinear programming are the most obvious examples of this. The convergence behavior of such methods is examined by comparing the sequence of iterates $\{x_k\}$ with a sequence $\{y_k\}$ of local minimizers for f in $\Omega(x_k)$. Issues of active constraint identification are also discussed in terms of the geometry of the sets $\Omega(x_k)$; conditions are given for x_{k+1} and y_k to lie on the same face of $\Omega(x_k)$.

Key words. constrained optimization, sequential quadratic programming, local convergence

AMS(MOS) subject classifications. 49D37, 90C30.

1. Introduction.

We consider the problem

$$\min f(x) \quad \text{s.t. } x \in \Omega \tag{1.1}$$

where $\Omega \subset \mathbb{R}^n$ and $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable. This problem has been analysed extensively by numerous authors (see, for example, the recent work of Dunn [5]-[7], Dunn and Sachs [8], Sachs [14], Calamai and Moré [3], and Burke and Moré [2]). Projected gradient, conditional gradient, and Newton-like methods have been proposed for its solution. In all these methods, it may be necessary at some stage to project a vector $x \in \mathbb{R}^n$ onto Ω, that is, to find

$$\arg\min \{\|x - z\| \mid z \in \Omega\}. \tag{1.2}$$

When Ω is geometrically simple (e.g., a disk, a cone, or a Cartesian product of these objects) or when Ω is defined by a set of linear equalities and inequalities, (1.2) may be computationally reasonable. However, when Ω is more complicated, it is impractical to repeatedly compute the projection (1.2). An example is when Ω is defined by a set of nonlinear equalities and inequalities, that is,

$$\Omega = \{z \mid c_j(z) \geqq 0, j = 1, \cdots, m_I, e_i(z) = 0, i = 1, \cdots, m_E\}. \tag{1.3}$$

A popular approach for such problems is known as sequential quadratic programming (SQP). At each iterate x_k, a "local linear approximation" $\Omega(x_k)$ to Ω is formed. A quadratic function is then minimized over this simpler set to obtain the next iterate x_{k+1}. This approach appears to have been originally suggested by Wilson [17].

The local convergence behavior of such methods is fairly well understood when "nondegeneracy" and "strict complementarity" conditions are satisfied at the solution x^*. Nondegeneracy conventionally refers to linear independence of the active constraint gradients at x^*, namely,

$$\{\nabla e_i(x^*), i = 1, \cdots, m_E, \nabla c_j(x^*), j \in \mathcal{A}^*\} \tag{1.4a}$$

where

$$\mathcal{A}^* = \{j \mid c_j(x^*) = 0\} \subset \{1, \cdots, m_I\}. \tag{1.4b}$$

* Received by the editors March 23, 1987; accepted for publication (in revised form) March 1, 1988.
† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

Strict complementarity means that the Lagrange multipliers satisfying the Kuhn-Tucker conditions, that is,

$$\nabla f(x^*) = \sum_{i=1}^{m_E} \lambda_i^E \nabla e_i(x^*) + \sum_{j \in \alpha^*} \lambda_j^I \nabla c_i(x^*), \tag{1.5}$$

are nonzero (in the case of the λ_i^E) and strictly positive (in the case of the λ_j^I). In this paper, we are concerned with local convergence when these assumptions are relaxed. In particular, when Ω is defined by (1.3), it is *not* assumed that the active constraint gradients are independent at x^*. A weaker nondegeneracy condition due to Dunn [7] is assumed instead. We examine the convergence of the sequence of iterates $\{x_k\}$ relative to a sequence $\{y_k\}$, where each y_k is a minimizer of $f(x)$ on the set $\Omega(x_k)$. (Of course, y_k is not computed.) Results concerning the convergence of $\{x_k\}$ to x^* are not obtained directly—these depend on the rate of convergence of $\{y_k\}$ to x^*, which may be linear when the constraint gradients are linearly dependent.

In § 2, we restate some definitions, due to Burke and Moré [2] and Clarke [4], which describe the geometry of sets in $\mathbb{R}^n$, the tangent and normal cones for such sets, and special subsets known as *faces*. The basic quadratic-programming subproblem is defined in § 3. Also, since the sequence of linear approximations $\{\Omega(x_k)\}$ does not generally converge to Ω, it may be that the SQP method will not converge to a solution of (1.1). Some sufficient conditions that ensure the limit point x^* *does* solve (1.1) are also discussed in § 3. In § 4, two theorems relating to the convergence are presented. Each uses a different approach, and makes different assumptions about the sequence of minimizers $\{y_k\}$ and the behavior of f in a feasible neighbourhood of the solution. Finally a discussion of possible choices of $\Omega(x_k)$ appears in § 5.

2. Definitions and notation. Here we restate some of the definitions of Clarke [4] and Burke and Moré [2], particularly those concerning faces of a convex set in $\mathbb{R}^n$.

For a general set $\Omega \subset \mathbb{R}^n$, the *tangent cone* $T(x; \Omega)$ at a point x is defined as "the set of vectors $u \in \mathbb{R}^n$ such that for every sequence w_i in Ω converging to x and every sequence t_i in $(0, \infty)$ converging to 0 there is a sequence u_i converging to u such that $w_i + t_i u_i \in \Omega$ for all i" (Clarke [4]). When Ω is *convex* this definition is equivalent to "the set of vectors $u \in \mathbb{R}^n$ such that there is a sequence w_i in Ω such that $(w_i - x)/\|w_i - x\|$ converges to $u/\|u\|$." The *normal cone* can be defined by polarity:

$$N(x; \Omega) = T^0(x; \Omega) = \{v \,|\, \langle u, v\rangle \leqq 0 \ \forall u \in T(x; \Omega)\}.$$

In the case of Ω convex,

$$N(x; \Omega) = \{v \,|\, \langle v, z - x\rangle \leqq 0 \ \forall z \in \Omega\}.$$

The projection operator P_Ω relative to Ω is

$$P_\Omega(x) = \arg\min \{\|z - x\| \,|\, z \in \Omega\}.$$

Clearly this is a contraction operator when Ω is convex, that is,

$$\|P_\Omega(x_1) - P_\Omega(x_2)\| \leqq \|x_1 - x_2\| \quad \forall x_1, x_2 \in \mathbb{R}^n$$

(see Calamai and Moré [3]).

The *affine hull* aff (S) of a set $S \subset \mathbb{R}^n$ is the smallest affine set that contains S, and we use ri (S) to denote the interior of S relative to aff (S). Using these definitions, we can state the first-order necessary conditions for a point $x^* \in \Omega$ to be optimal in problem (1.1).

DEFINITION 2.1. $x^* \in \Omega$ is said to be a *stationary point* for (1.1) if

$$-\nabla f(x^*) \in N(x^*; \Omega). \tag{2.1}$$

THEOREM 2.2 (Clarke [4]). *Condition* (2.1) *is a first-order necessary condition for* x^* *to be optimal for* (1.1).

Dunn's [7] nondegeneracy condition then follows.

DEFINITION 2.3. A stationary point x^* is said to be *nondegenerate* if

$$-\nabla f(x^*) \in \text{ri}\,(N(x^*; \Omega)).$$

Consider now a *convex* set $\Omega_* \subset \mathbb{R}^n$. There is a special class of subsets of Ω_* known as *faces*, that can be defined as follows (see Burke and Moré [2, Thm. 2.1], Rockafellar [12]).

DEFINITION 2.4. A convex subset Ω_F of Ω_* is a *face* of Ω_* if, for all convex $\Gamma \subset \Omega_*$ such that ri (Γ) meets Ω_F, we have $\Gamma \subset \Omega_F$.

A face $\Omega_E \subset \Omega_*$ is said to be *exposed* if

$$\Omega_E = \arg\max \{\phi(x) \,|\, x \in \Omega_*\}$$

where ϕ is some linear functional. It is proved in Burke and Moré [2] that the normal and tangent cones to Ω are the same for all $x \in \text{ri}\,(\Omega_F)$; hence we can use the notation $N(\Omega_F; \Omega_*)$ instead of $N(x; \Omega_*)$, where $x \in \text{ri}\,(\Omega_F)$.

A face Ω_F is referred to as a *quasi-polyhedral* face of Ω_* if

$$\text{aff}\,(\Omega_F) = x + \text{lin}\,(T(x)),$$

for any $x \in \Omega_F$, where the lineality lin $(T(x))$ is defined by

$$\text{lin}\,(T(x)) = T(x) \cap (-T(x)).$$

Some examples of such faces are given by Burke and Moré. In the special case of Ω_* polyhedral, all faces are quasi-polyhedral. The definition above is closely related to that of an *open facet*, as proposed by Dunn [7].

An important result regarding quasi-polyhedral faces is proved in [2].

THEOREM 2.5 [2, Thm. 2.8]. *Let* Ω_F *be a nonempty face of a convex set* Ω_*. *Then* Ω_F *is a nonempty quasi-polyhedral face if and only if* $\Omega_F + N(\Omega_F; \Omega_*)$ *has an interior. When this is true,*

$$\text{int}\,(\Omega_F + N(\Omega_F; \Omega_*)) = \text{ri}\,(\Omega_F) + \text{ri}\,(N(\Omega_F; \Omega_*)).$$

If Ω_F is a quasi-polyhedral face such that $x^* \in \text{ri}\,(\Omega_F)$, and if x^* is a nondegenerate stationary point, that is, $-\nabla f(x^*) \in \text{ri}\,(N(\Omega_F; \Omega_*))$, then it follows from Theorem 2.5 that

$$x^* - \nabla f(x^*) \in \text{int}\,(\Omega_F + N(\Omega_F; \Omega_*)).$$

This observation is used by Burke and Moré to obtain results concerning the identification of active constraints, and will be used here in subsequent sections.

Throughout this paper we use B to denote the closed unit ball in $\mathbb{R}^n$ and $\|\cdot\|$ to denote the Euclidean norm.

3. SQP methods—optimality conditions and active constraint identification. In this section we consider sequential quadratic programming methods for solving (1.1), in which the sequence $\{x_k\}$ is generated according to the following scheme:

$$\begin{aligned} &\text{At } x_k, \text{ solve} \\ &\min_{p_k} \langle p_k, \nabla f(x_k)\rangle + \tfrac{1}{2}\langle p_k, B_k p_k\rangle \\ &\text{s.t.} \quad x_k + p_k \in \Omega(x_k); \\ &\text{set} \quad x_{k+1} = x_k + p_k. \end{aligned} \tag{3.1}$$

Here $\{B_k\}$ is a bounded sequence of symmetric matrices and $\Omega(\cdot)$ can be regarded as

a multifunction that maps $\mathbb{R}^n$ to subsets of $\mathbb{R}^n$. Previous work by Dunn [5], [6], Burke and Moré [2], Sachs [14], and others has dealt with the case $\Omega(x_k) = \Omega$. That is, all iterates x_k are feasible with respect to the original feasible set Ω. If, as is usual in these papers, Ω is assumed to be convex, then $x_k + \lambda p_k \in \Omega$ for all $\lambda \in [0, 1]$, and so a line search can be used in such a way as to ensure global convergence of the method. This is not appropriate in (3.1), as it is possible that $x_k + \lambda p_k \notin \Omega(x_k)$ for all $\lambda \in [0, 1)$. Hence we are only concerned here with issues of local convergence, and we assume that the full step p_k is taken at each iteration.

In the remainder of the paper we use the following assumptions.

Assumption 3.1. (i) $\Omega(x)$ is closed and convex for all $x \in \mathbb{R}^n$.

(ii) If $\{x_k\}$ converges to x^*, then $\Omega(x_k)$ converges in the Kuratowski sense to some convex set $\Omega_* \subset \mathbb{R}^n$. That is,

(a) All sequences $\{y_k\}$ with $y_k \in \Omega(x_k)$ have all their accumulation points in Ω_*;

(b) For all $y \in \Omega_*$ there is a sequence $\{y_k\}$ with $y_k \in \Omega(x_k)$ such that $\lim y_k = y$.

Assumption 3.2. Algorithm (3.1) generates a sequence $\{x_k\}$ that converges to a point $x^* \in \Omega$ which is a nondegenerate stationary point of f in Ω_*. (Note that $x^* \in \Omega_*$ follows from (3.1) and Assumption 3.1.)

Notes. (i) Kuratowski convergence of sequences of sets is discussed in more detail in Salinetti and Wets [15] and Attouch [1]. For further information on convex multifunctions of the form $\Omega(x_k)$, see Robinson [11] and Rockafellar [13]. Assumption 3.1(ii) is weaker than any of the variants of Lipschitz continuity of multifunctions discussed in Rockafellar [13]. Note also that we do *not* assume above that $\Omega_* = \Omega(x^*)$.

(ii) Assumptions of the form 3.2 are usually made when the *local* convergence properties of a method are being studied. The purpose of a *global* convergence analysis is to show that convergence to a stationary point occurs from any given starting point. Since substantial modifications are usually required to ensure global convergence of SQP methods, we do not perform a global analysis here.

We do not assume that x^* is stationary with respect to Ω, and the following example shows that this does not generally follow from Assumption 3.2.

Example 3.3. Consider

$$\min x \quad \text{s.t. } c(x) = x^2 \geqq 0.$$

Applying the usual SQP algorithm starting at the point $x_0 = 1$, with $B_k \equiv 0$ and

$$\begin{aligned}\Omega(x_k) &= \{z \mid \nabla c(x_k)^T (z - x_k) + c(x_k) \geqq 0\} \\ &= \{z \mid 2x_k z - x_k^2 \geqq 0\},\end{aligned}$$

we obtain the sequence $x_k = 2^{-k}$ which converges to $x^* = 0$. For this problem $\Omega = \Omega(0) = \mathbb{R}$, but $\Omega_* = \mathbb{R}^+$. The point x^* is optimal in Ω_* but is not even a stationary point in Ω.

The following theorem gives a sufficient condition for x^* to be a stationary point of (1.1).

THEOREM 3.4. *Suppose x^* is a stationary point of f in Ω_*. If $T(x^*; \Omega) \subset T(x^*; \Omega_*)$ then x^* is also a stationary point of f in Ω.*

Proof. The proof follows from $N(x^*; \Omega_*) \subset N(x^*; \Omega)$. □

To find situations in which the condition of Theorem 3.4 holds, consider the nonlinear programming problem in its standard form:

$$\begin{aligned}&\min f(x) \\ &\text{s.t.} \quad c_j(x) \geqq 0, \qquad j = 1, \cdots, m_I, \\ &\qquad\quad e_i(x) = 0, \qquad i = 1, \cdots, m_E.\end{aligned} \tag{3.2}$$

Here the feasible set Ω is given by (1.3). The functions f, c_j, and e_i are assumed to be twice continuously differentiable. The multifunction $\Omega(x)$ is usually generated by linearization of the constraints about x:

$$\Omega(x) = \Lambda(x) \overset{\text{def}}{=} \left\{ z \,\middle|\, \begin{array}{ll} \langle z - x, \nabla c_j(x)\rangle + c_j(x) \geqq 0, & j = 1, \cdots, m_I \\ \langle z - x, \nabla e_i(x)\rangle + e_i(x) = 0, & i = 1, \cdots, m_E \end{array} \right\}. \tag{3.3}$$

An often-used sufficient condition for a stationary point x^* of f in Ω to be a Kuhn-Tucker point for problem (3.2) is that

$$T(x^*; \Omega) = T(x^*; \Lambda(x^*)) = \left\{ u \,\middle|\, \begin{array}{ll} \langle u, \nabla c_j(x^*)\rangle \geqq 0, & j \in \mathscr{A}^* \\ \langle u, \nabla e_i(x^*)\rangle = 0, & i = 1, \cdots, m_E \end{array} \right\} \tag{3.4}$$

where $\mathscr{A}^*$ is defined in (1.4b). This is referred to as the *Guignard constraint qualification* (see Gould and Tolle [10] and Fletcher [9]).

The conditions $T(x^*; \Omega) \subset T(x^*; \Omega_*)$ and (3.4) are not equivalent, nor does one imply the other. In Example 3.3, clearly $T(0; \Omega_*) = \mathbb{R}^+$ and $T(0; \Omega) = \mathbb{R}$, but $\Omega(0) = \Lambda(0) = \mathbb{R}$, and so (3.4) is satisfied. On the other hand, we have in the following example that $T(x^*; \Omega) \subset T(x^*; \Omega_*)$, but (3.4) is not satisfied.

Example 3.5. The problem

$$\min_{z \in \mathbb{R}^2} z_1 \quad \text{s.t. } z_2 \geqq 0, \quad z_2 \leqq z_1^3$$

has solution $x^* = 0$. From (3.3)

$$\Lambda(z) = \{y \,|\, y_2 \geqq 0,\ y_2 \leqq 3z_1^2 y_1 - 2z_1^3\}.$$

If the starting point $x_0 = (1, 0)^T$ is used, the method (3.1) with $B_k \equiv 0$ generates the sequence $x_k = ((\frac{2}{3})^k, 0)^T$. Clearly $\Omega_* = \{(y_1, 0)^T \,|\, y_1 \in \mathbb{R}^+\}$ and $T(0; \Omega_*) = \Omega_*$. However,

$$\Lambda(0) = \{(y_1, 0)^T \,|\, y_1 \in \mathbb{R}\},$$

and so condition (3.4) is not satisfied.

The following result gives sufficient conditions for both $T(x^*; \Omega_*) = T(x^*; \Omega)$ and (3.4), in some familiar cases.

THEOREM 3.6. *Consider problem* (3.2) *and assume that the* c_j, e_i *are all twice continuously differentiable. Suppose* $x^* \in \Omega$, *and let* $\{x_k\}$ *be any sequence converging to* x^*. *Assume also that* $\Omega(x) = \Lambda(x)$ (*from* (3.3)). *Then sufficient conditions for* $T(x^*; \Omega) = T(x^*; \Omega_*)$, *and condition* (3.4), *are the following*:

(a) $c_j(x)$, $j = 1, \cdots, m_I$ *and* $e_i(x)$, $i = 1, \cdots, m_E$ *are linear functions*;

(b) *The set* $\{\nabla e_i(x^*), i = 1, \cdots, m_E, \nabla c_j(x^*), j \in \mathscr{A}^*\}$ *is linearly independent.*

Proof. (a) The proof follows from the fact that $\Lambda(x) = \Omega = \Omega_*$ for all x, and since Ω is polyhedral, (3.4) holds for all feasible x^*.

(b) The inclusions $T(x^*; \Omega) \subset T(x^*; \Lambda(x^*))$ and $T(x^*; \Omega_*) \subset T(x^*; \Lambda(x^*))$ follow from the assumed continuity properties of the c_j and e_i, and Assumption 3.1. Linear independence is not needed here. The reverse inclusion $T(x^*; \Lambda(x^*)) \subset T(x^*; \Omega)$ is proved in Fletcher [9, Lemma 9.2.2] using the linear independence assumption.

We complete the proof by showing that $T(x^*; \Omega) \subset T(x^*; \Omega_*)$. Choose $u \in T(x^*; \Omega) \backslash \{0\}$ and assume without loss of generality (since $T(x^*; \Omega)$ is a cone) that $\|u\| = 1$. Then there are sequences $u_m \to u$ and $t_m \downarrow 0$ such that

$$\begin{aligned} c_j(x^* + t_m u_m) &\geqq 0, \qquad j = 1, \cdots, m_I, \\ e_i(x^* + t_m u_m) &= 0, \qquad i = 1, \cdots, m_E. \end{aligned}$$

Hence

$$
\begin{aligned}
&c_j(x_k)+\langle \nabla c_j(x_k),\, x^*+t_m u_m-x_k\rangle + d_j^{(1)} \geqq 0,\\
&e_i(x_k)+\langle \nabla e_i(x_k),\, x^*+t_m u_m-x_k\rangle + d_i^{(2)} = 0
\end{aligned} \tag{3.5}
$$

where $\|d_j^{(1)}\| = O(\|x^*+t_m u_m-x_k\|^2) = \|d_i^{(2)}\|$. Now it follows from the linear independence assumption and the fact that $x_k \to x^*$ that $\nabla c_j(x_k)$, $j \in \mathscr{A}^*$ and $\nabla e_i(x_k)$, $i = 1, \cdots, m_E$ are linearly independent for k sufficiently large. Hence for such k there is a vector g_{km} that satisfies

$$
\begin{aligned}
&\langle \nabla c_j(x_k), g_{km}\rangle = d_j^{(1)}, \qquad j \in \mathscr{A}^*,\\
&\langle \nabla e_i(x_k), g_{km}\rangle = d_i^{(2)}, \qquad i = 1, \cdots, m_E,\\
&g_{km} = O(\|x_k+t_m u_m-x^*\|^2).
\end{aligned}
$$

Hence from (3.5),

$$
\begin{aligned}
&c_j(x_k)+\langle \nabla c_j(x_k),\, x^*+t_m u_m+g_{km}-x_k\rangle \geqq 0, \quad j \in \mathscr{A}^*,\\
&e_i(x_k)+\langle \nabla e_i(x_k),\, x^*+t_m u_m+g_{km}-x_k\rangle = 0, \quad i = 1, \cdots, m_E.
\end{aligned} \tag{3.6}
$$

Note that the inactive inequalities in (3.5) will always be satisfied for all k, m sufficiently large. Now from (3.6) we have that

$$
x^*+t_m u_m+g_{km} \in \Lambda(x_k),
$$

so choosing a subsequence in k if necessary and taking the limit as $k \to \infty$ we obtain by Assumption 3.1 that

$$
x^*+t_m u_m+\bar{g}_m \in \Omega_* \tag{3.7}
$$

where $\bar{g}_m = \lim_k g_{km} = O(t_m^2)$. Defining

$$
\bar{u}_m = u_m + t_m^{-1}\bar{g}_m,
$$

we have from (3.7) that $x^*+t_m\bar{u}_m \in \Omega_*$, and that $\bar{u}_m \to u$, $t_m \downarrow 0$. Hence $u \in T(x^*; \Omega_*)$. □

The following result discusses a less familiar situation in which the condition of Theorem 3.4 holds. We make use of functions h_j that provide general measures of the curvature of the constraint functions near x^*. The result shows that when Ω is an intersection of convex sets, at least locally, then $T(x^*; \Omega) \subset T(x^*; \Omega_*)$.

THEOREM 3.7. *Suppose Ω is defined by* (1.3) *with only inequality constraints* ($m_E = 0$). *Let $x^* \in \Omega$ be such that each c_j is twice continuously differentiable at x^*. For each active constraint $j \in \mathscr{A}^*$ define the function*

$$
h_j(v) = c_j(x^*+v) - \langle \nabla c_j(x^*), v\rangle,
$$

and the set

$$
N_j = \{u \mid \|u\| = 1,\ u \in T(x^*; \Omega) \text{ and } \langle u, \nabla c_j(x^*)\rangle = 0\}.
$$

Assume that, for each $j \in \mathscr{A}^*$, either c_j is a linear function or there exist constants $\delta > 0$ and $T > 0$ such that if N_j is not empty,

$$
h_j(tv) < 0 \quad \text{for all } t \in (0, T], \quad v \in N_j + \delta B. \tag{3.8}
$$

Then if $\{x_k\}$ is any sequence converging to x^*, and $\Omega(x) = \Lambda(x)$, then $T(x^*; \Omega) \subset T(x^*; \Omega_*)$.

Proof. Take $u \in T(x^*; \Omega)\backslash\{0\}$ and assume without loss of generality that $\|u\| = 1$. Then there are sequences $u_m \to u$, $t_m \downarrow 0$ such that $x^* + t_m u_m \in \Omega$, that is,

$$c_j(x^* + t_m u_m) \geqq 0, \qquad j = 1, \cdots, m_I.$$

We aim to show that $x^* + t_m u_m \in \Lambda(x_k)$ for m sufficiently large, and for $k \geqq k_m$, where k_m is a positive integer to be defined. That is,

$$c_j(x_k) + \langle \nabla c_j(x_k), x^* + t_m u_m - x_k \rangle \geqq 0, \qquad j = 1, \cdots, m_I. \tag{3.9}$$

This will ensure that $x^* + t_m u_m \in \Omega_*$. Clearly (3.9) will be satisfied by the linear constraints, and will eventually be satisfied by the inactive constraints $j \notin \mathcal{A}^*$. For $j \in \mathcal{A}^*$ we have

$$\begin{aligned} & c_j(x^* + t_m u_m) \geqq 0, \\ \Leftrightarrow\ & \langle \nabla c_j(x^*), t_m u_m \rangle \geqq -h_j(t_m u_m), \\ \Leftrightarrow\ & c_j(x_k) + \langle \nabla c_j(x_k), t_m u_m \rangle \geqq -h_j(t_m u_m) + c_j(x_k) + \langle \nabla c_j(x_k) - \nabla c_j(x^*), t_m u_m \rangle. \end{aligned} \tag{3.10}$$

By taking the limit in k in (3.10) and noting that $h(t_m u_m) = O(t_m^2)$, we obtain

$$\langle \nabla c_j(x^*), u \rangle \geqq 0.$$

If this inequality is strict then $\langle \nabla c_j(x^*), u_m \rangle > 0$ for sufficiently large m, and then k_m can be chosen large enough to ensure (3.9). Otherwise $u \in N_j$, and so by the assumption of the theorem $h_j(t_m u_m) < 0$ for m sufficiently large. If we choose k_m large enough so that

$$\begin{aligned} & \left| c_j(x_k) + \langle \nabla c_j(x_k) - \nabla c_j(x^*), t_m u_m \rangle \right| \leqq -\tfrac{1}{2} h_j(t_m u_m), \\ & \left| \langle \nabla c_j(x_k), x^* - x_k \rangle \right| \leqq -\tfrac{1}{2} h_j(t_m u_m) \end{aligned}$$

for all $k \geqq k_m$, (3.9) follows from (3.10). □

Example 3.8. Let

$$\Omega = \{x \in \mathbb{R}^2 \,|\, c_1(x) = x_2 - x_1^K \geqq 0,\ c_2(x) = x_2 + x_1 \geqq 0,\ c_3(x) = x_2 \geqq 0\}$$

where K is some positive integer, at least 2, and let $x^* = (0, 0)^T$. Then

$$\Lambda(x) = \{y \in \mathbb{R}^2 \,|\, y_2 \geqq K x_1^{K-1} y_1 - (K-1) x_1^K,\ y_2 + y_1 \geqq 0,\ y_2 \geqq 0\}$$

and clearly, for any sequence converging to x^*,

$$\begin{aligned} & \Omega_* = \Lambda(x^*) = \{y \in \mathbb{R}^2 \,|\, y_2 + y_1 \geqq 0,\ y_2 \geqq 0\}, \\ & T(x^*; \Omega) = T(x^*; \Omega_*). \end{aligned}$$

By the definitions of Theorem 3.7,

$$N_1 = \{(1, 0)^T\} \quad \text{and} \quad h_1(tv) = -t^K v_1^K,$$

and condition (3.8) can be satisfied by choosing any δ with $0 < \delta < 1$, and any $T > 0$. Note that the conditions of Theorem 3.6 do not apply.

Note that a more intuitive, but more restrictive, condition than (3.8) would be that

$$\langle u, \nabla^2 c_j(x^*) u \rangle \leqq -\alpha_j < 0 \quad \text{for all } u \in N_j.$$

This condition implies (3.8) since

$$\begin{aligned} h(tv) &= \tfrac{1}{2} t^2 \langle v, \nabla^2 c_j(x^*) v \rangle + O(t^3) \\ &\leqq -\tfrac{1}{4} t^2 \alpha_j \leqq 0 \end{aligned}$$

for δ, T sufficiently small, and $t \in [0, T]$.

In the case in which Ω is defined by (1.3), the issue of active set identification consists of finding those indices $j = 1, \cdots, m_I$ for which $c_j(x^*) = 0$, before x^* itself has been identified exactly. This is usually done by checking the sign of the Lagrange multipliers at each iterate x_k, and by finding the linearized constraints from the previous iteration with respect to which x_k is active. Using the general formulation (3.1), we can instead state this issue in terms of finding the face of Ω_* on which x^* lies, and determining whether this face can be "predicted" by the face of $\Omega(x_k)$ on which x_{k+1} lies.

In the subsequent discussion, let Ω_F be the face of Ω_* for which $x^* \in \text{ri}\,(\Omega_F)$. The next results show how the faces of $\Omega(x_k)$ are related to Ω_F.

THEOREM 3.9. *Let $\{x_k\}$ be a sequence of points converging to x^*, with $x_{k+1} \in \Omega(x_k)$, and suppose that the sequence $\{\Omega(x_k)\}$ satisfies Assumptions 3.1. Let Ω_F be the face of Ω_* for which $x^* \in \text{ri}\,(\Omega_F)$. Suppose $\{y_k\}$ and $\{\nu_k\}$ are two sequences such that $y_k \in \Omega(x_k)$. $\nu_k \in N(y_k; \Omega(x_k))$, with $\nu_k \to \nu^* \in N(\Omega_F; \Omega_*)$. Then any accumulation point y^* of $\{y_k\}$ satisfies*

$$y^* \in \Omega_E \overset{\text{def}}{=} \arg\max \{\langle \nu^*, x\rangle \,|\, x \in \Omega_*\} \quad \textit{and} \quad \Omega_F \subset \Omega_E.$$

Proof. Clearly, by the assumptions on ν^* and x^*, the last part of the theorem is true.

By Assumptions 3.1, $y^* \in \Omega_*$. Assume without loss of generality that $y_k \to y^*$. If $y^* \notin \Omega_E$, then there is some point $z \in \Omega_*$ such that

$$\langle \nu^*, z - y^*\rangle > 0.$$

Defining the sequence $z_k = P_{\Omega(x_k)}(z)$, we have that $z_k \to z$, and hence that $\langle \nu^*, z_k - y^*\rangle > 0$ for k sufficiently large. In fact, by continuity, $\langle \nu_k, z_k - y_k\rangle > 0$ for k sufficiently large, which contradicts the assumption that $\nu_k \in N(y_k; \Omega(x_k))$. □

The inclusion $\Omega_F \subset \Omega_E$ can be strict unless Ω_F is quasi-polyhedral and $\nu^* \in \text{ri}\,(N(\Omega_F; \Omega_*))$, as the following result shows.

THEOREM 3.10. *Suppose there is a nonempty quasi-polyhedral face Ω_F of the convex set Ω_* and an element $\nu^* \in N(\Omega_F; \Omega_*)$. Define the face Ω_E as in Theorem 3.7. If $\nu^* \in \text{ri}\,(N(\Omega_F; \Omega_*))$, then $\Omega_E = \Omega_F$; otherwise it is possible that the inclusion $\Omega_F \subset \Omega_E$ is strict.*

Proof. Assume $\nu^* \in \text{ri}\,(N(\Omega_F; \Omega_*))$. We need only prove the inclusion $\Omega_E \subset \Omega_F$. Choosing points $x_0 \in \text{ri}\,(\Omega_F) \subset \Omega_E$ and $x_1 \in \Omega_E$, we have for all $\lambda \in [0, 1]$ that

$$x_\lambda = (1 - \lambda)x_0 + \lambda x_1 \in \Omega_E.$$

Also $\nu^* \in N(\Omega_E; \Omega_*)$, and so

$$P_{\Omega_*}(x_\lambda + \nu^*) = x_\lambda.$$

However, since $x_0 + \nu^* \in \text{int}\,(\Omega_F + N(\Omega_F; \Omega_*))$ (by Theorem 2.8 of Burke and Moré [2]) we have for some small positive λ that

$$x_\lambda + \nu^* \in \text{int}\,(\Omega_F + N(\Omega_F; \Omega_*)).$$

Hence

$$x_\lambda = P_{\Omega_*}(x_\lambda + \nu^*) \in \text{ri}\,(\Omega_F),$$

and so by Definition 2.4, $x_1 \in \Omega_F$, giving the first result.

For the second part, consider the example

$$\Omega_* = \{x \in \mathbb{R}^2 \,|\, x_2 \geqq |x_1|\}, \quad \Omega_F = (0, 0)^T, \quad \nu^* = (1, -1)^T.$$

Then

$$\Omega_E = \{x \in \mathbb{R}^2 \,|\, x_1 \geqq 0,\ x_1 = x_2\}.$$

□

COROLLARY 3.11. *Suppose the assumptions of Theorem* 3.10 *are satisfied, and that* $\{y_k\}$ *and* $\{v_k\}$ *are defined as in Theorem* 3.9 *with* $v^* \in \text{ri}\,(N(\Omega_F; \Omega_*))$. *Then all accumulation points of the sequence* $\{y_k\}$ *lie on* Ω_F. □

The assumption that Ω_F is quasi-polyhedral certainly holds when $\Omega(x_k) = \Lambda(x_k)$ (see (3.3)). Then each $\Omega(x_k)$ is a polyhedron, and hence Ω_* is also polyhedral. In this case *all* faces of Ω_* are quasi-polyhedral.

4. Convergence of SQP methods. Some results concerning the convergence of algorithm (3.1) are presented in this section. For the case of standard nonlinear programming (3.2), the convergence behavior is well known when the active constraint gradients are linearly independent at x^*. Here we establish results for the convergence of $\{x_k\}$ when this is not necessarily true.

It is assumed throughout that x^* is an *isolated local minimizer* of f in Ω_*, that is, there is $R > 0$ such that

$$f(x^*) < f(x) \quad \text{for all } x \in \Omega_* \cap (x^* + RB). \tag{4.1}$$

Convergence results are obtained by comparing the sequence $\{x_k\}$ with a sequence $\{y_k\}$ consisting of local minimizers of f in $\Omega(x_k)$. Lemma 4.1 shows that it is possible to choose the sequence $\{y_k\}$ such that $y_k \to x^*$.

The result makes use of the concept of epiconvergence of a sequence of functions, as discussed by Attouch [1, p. 26]. Briefly, we say that a sequence of functions $\{F_k\}$ is epiconvergent if for all $x \in \mathbb{R}^n$,

$$\sup_{V \in \mathcal{N}(x)} \liminf_k \inf_{u \in V} F_k(u) = \sup_{V \in \mathcal{N}(x)} \limsup_k \inf_{u \in V} F_k(u)$$

where $\mathcal{N}(x)$ is the set of all open neighbourhoods of x in $\mathbb{R}^n$. For $\varepsilon > 0$ we need to define the set

$$\varepsilon - \arg\min G = \{x \mid G(x) \leqq -1/\varepsilon \text{ or } G(x) \leqq \inf_{u \in \mathbb{R}^n} G(u) + \varepsilon\}.$$

The lim sup of a sequence of sets $\{S_k\}$ is defined similarly to Assumption 3.1(ii)(b), that is,

$$\bar{S} = \limsup_k S_k$$

if each $\bar{x} \in \bar{S}$ is an accumulation point of a sequence $\{x_k\}$ with $x_k \in S_k$. Finally, define the indicator function of a set S as

$$\Psi(x \mid S) = \begin{cases} 0, & x \in S, \\ +\infty, & x \notin S. \end{cases}$$

LEMMA 4.1. *Suppose Assumptions* 3.1 *and* 3.2 *are satisfied, and* x^* *is an isolated local minimizer of some twice continuously differentiable function* f *in* Ω_*. *Then there exists a sequence* y_k *such that* $y_k \in \Omega(x_k)$ *is a local minimizer of* f *in* $\Omega(x_k)$, *and* $y_k \to x^*$.

Proof. Let $R \geqq 0$ be chosen to satisfy (4.1). Define a sequence of functions

$$F_k(x) = f(x) + \Psi(x \mid \Omega_R(x_k)) \quad \text{where } \Omega_R(x_k) = \Omega(x_k) \cap (x^* + RB).$$

By Theorem 1.39 of Attouch [1], Kuratowski convergence of the sequence $\{\Omega_R(x_k)\}$ (which follows from our Assumptions 3.1) implies epiconvergence of the corresponding indicator functions. Hence the sequence $\{F_k\}$ is epiconvergent, with a limiting function

$$F_*(x) = f(x) + \Psi(x \mid \Omega_{R^*}) \quad \text{with } \Omega_{R^*} = \Omega_* \cap (x^* + RB).$$

Now Proposition 2.9 of Attouch [1] states that for any positive sequence $\varepsilon_k \downarrow 0$,

$$\limsup_k (\varepsilon_k - \arg\min F_k) \subset \arg\min F_*. \tag{4.2}$$

Since $\Omega_R(x_k)$ is closed, F_k takes on its minimum value at some point $y_k \in \Omega_R(x_R)$, and

$$y_k \in \varepsilon_k - \arg\min F_k.$$

Since $\arg\min F_* = \{x^*\}$, then by (4.2) all accumulation points of $\{y_k\}$ are x^*, and since by definition $y_k \in x^* + RB$ (a closed set) for all k, we have

$$\lim y_k = x^*,$$

as required. □

Each iteration of (3.1) entails the minimization of a quadratic approximation of $f(x)$ over a convex set $\Omega(x_k)$. Since the true minimizer of f on this set is y_k, it seems appropriate to establish a relationship between $\|x_k - y_k\|$ and $\|x_{k+1} - y_k\|$.

THEOREM 4.2. *Suppose $\Omega(x_k) \subset \mathbb{R}^n$ is a convex set and y_k is a local minimizer of a twice continuously differentiable function f in $\Omega(x_k)$. Let x_{k+1} be generated by the algorithm* (3.1). *Suppose Ω_{kF} is the face of $\Omega(x_k)$ for which $y_k \in \operatorname{ri}(\Omega_{kF})$, and that Ω_{kF} is quasi-polyhedral. Then there is a small positive number ε_k such that if*

$$\langle \mu, B_k \mu \rangle \geqq a_k \langle \mu, \mu \rangle \quad \text{for all } \mu \in \operatorname{lin}(T(\Omega_{kF}; \Omega(x_k))) \text{ and some } a_k > 0, \tag{4.3}$$

and $\|x_k - y_k\| \leqq \varepsilon_k$, then $x_{k+1} \in \operatorname{ri}(\Omega_{kF})$.

(*Note.* This result is similar to Theorem 4.1 of Burke and Moré [2]. However, we state a proof below that will be useful in subsequent discussions.)

Proof. We first seek a solution p_k of (3.1) such that $x_{k+1} = x_k + p_k \in \operatorname{aff}(\Omega_{kF})$. Then it is shown that $\|p_k\| = O(\|x_k - y_k\|) = O(\varepsilon_k)$, and hence that $\|x_{k+1} - y_k\| = O(\varepsilon_k)$. Since $y_k \in \operatorname{ri}(\Omega_{kF})$, it then follows that $x_{k+1} \in \operatorname{ri}(\Omega_{kF})$ for ε_k sufficiently small.

Since we are assuming $x_k + p_k \in \operatorname{aff}(\Omega_{kF})$, and since $y_k \in \Omega_{kF} \subset \operatorname{aff}(\Omega_{kF})$, we have

$$\begin{aligned} &\langle x_k + p_k - y_k, \nu \rangle = 0, \\ \Leftrightarrow &\langle p_k, \nu \rangle = \langle y_k - x_k, \nu \rangle \quad \text{for all } \nu \in N(\Omega_{kF}; \Omega(x_k)). \end{aligned} \tag{4.4}$$

Using the optimality conditions for (3.1), we have

$$-\nabla f(x_k) - B_k p_k \in N(\Omega_{kF}; \Omega(x_k)),$$

and hence

$$\langle \nabla f(x_k) + B_k p_k, z_1 - y_k \rangle = 0$$

for all $z_1 \in \operatorname{aff}(\Omega_{kF})$. A similar equation holds for $\nabla f(y_k)$, by the optimality of y_k in $\Omega(x_k)$:

$$\langle \nabla f(y_k), z_1 - y_k \rangle = 0 \quad \text{for all } z_1 \in \operatorname{aff}(\Omega_{kF}).$$

Since Ω_{kF} is quasi-polyhedral,

$$\operatorname{aff}(\Omega_{kF}) = y_k + \operatorname{lin}(T(\Omega_{kF}; \Omega(x_k))),$$

and so, from the two previous equations,

$$\langle B_k p_k, \mu \rangle = \langle \nabla f(y_k) - \nabla f(x_k), \mu \rangle \quad \text{for all } \mu \in \operatorname{lin}(T(\Omega_{kF}; \Omega(x_k))). \tag{4.5}$$

From Theorem 2.5, the vectors ν and μ in (4.4) and (4.5) span $\mathbb{R}^n$. In addition, (4.3) implies that p_k is uniquely determined by (4.4) and (4.5), and from the right-hand sides of these equations we have that

$$\|p_k\| = O(\|x_k - y_k\|).$$

Note that the second-order sufficient conditions for p_k to be a minimizer of (3.1) are satisfied because of (4.3). So we have found $x_{k+1} = x_k + p_k$ such that $x_{k+1} \in \operatorname{aff}(\Omega_{kF})$ and $\|x_{k+1} - y_k\| = O(\|x_k - y_k\|) = O(\varepsilon_k)$. Since $y_k \in \operatorname{ri}(\Omega_{kF})$ and $\operatorname{ri}(\Omega_{kF})$ is the interior of Ω_{kF} relative to $\operatorname{aff}(\Omega_{kF})$, it follows that $x_{k+1} \in \operatorname{ri}(\Omega_{kF})$ for ε_k sufficiently small. □

In the proof above it is not necessary for y_k to be a *nondegenerate* minimizer in $\Omega(x_k)$, nor is it assumed that $x_k \in \Omega(x_k)$.

THEOREM 4.3. *Let f be twice continuously differentiable, and suppose that the sequence $\{x_k\}$ generated by* (3.1) *with $B_k = \nabla^2 f(x_k)$ converges to x^*, an isolated local minimizer of f in Ω_*. Suppose Assumptions* 3.1 *and* 3.2 *are satisfied, and let $\{y_k\}$ be chosen as in Lemma* 4.1, *so that $y_k \to x^*$.*

For some k sufficiently large, assume that the face Ω_{kF} of $\Omega(x_k)$ for which $y_k \in \text{ri}\,(\Omega_{kF})$ is quasi-polyhedral, and that

$$\langle \mu, \nabla^2 f(x^*)\mu\rangle \geqq b_k\langle \mu, \mu\rangle \quad \text{for some } b_k > 0, \text{ and all } \mu \in \text{lin}\,(T(\Omega_{kF}; \Omega(x_k))). \tag{4.6}$$

Then there is $\varepsilon_k > 0$ such that if

$$\|x_k - y_k\| \leqq \varepsilon_k,$$

then

$$x_{k+1} \in \text{ri}\,(\Omega_{kF}) \quad \text{and} \quad \|x_{k+1} - y_k\| = O(\|x_k - y_k\|^2).$$

Proof. Set $a_k = \frac{1}{2}b_k$ and $B_k = \nabla^2 f(x_k)$. Then for k sufficiently large, (4.6) implies (4.3). Hence from Theorem 4.2 there is $\varepsilon_k > 0$ such that $x_k + p_k \in \Omega_{kF}$. Further, from (4.5) with $B_k = \nabla^2 f(x_k)$, we find that

$$\langle \nabla^2 f(x_k)(x_k + p_k - y_k), \mu\rangle = O(\|x_k - y_k\|^2) \quad \text{for all } \mu \in \text{lin}\,(T(\Omega_{kF}; \Omega(x_k))).$$

Combining this with (4.4), we obtain

$$\|x_k + p_k - y_k\| = O(\|x_k - y_k\|^2). \qquad \square$$

The result above follows from the analysis of Newton's method on a convex set (see, for example, Dunn [6], Dunn and Sachs [8], Sachs [14], and the references cited in these sources). It is shown in those papers that conditions other than (4.6) can be used to obtain convergence results. Below, the analysis of Sachs [14] is used to prove a variant of Theorem 4.3 in which the assumption of quasi-polyhedrality of Ω_{kF} is not required, and an alternative to (4.6) is used. We show subsequently that the alternative condition is neither weaker nor stronger than (4.6).

We start by defining a function that gives a measure of the increase of a quadratic approximation to f in a feasible neighbourhood of a given point ξ. Let V be a closed bounded set in $\mathbb{R}^n$ and suppose $\xi \in V$. Define

$$c(\sigma; \xi, V) = \inf_{\substack{\|x-\xi\| \geqq \sigma \\ x \in V}} \langle x - \xi, \nabla f(\xi)\rangle + \tfrac{1}{2}\langle x - \xi, \nabla^2 f(\xi)(x - \xi)\rangle.$$

In the following result we make use of the sets $\Omega_R(x_k)$ and Ω_{R^*}, as defined in the proof of Lemma 4.1, and show how the function $c(\sigma; x^*, \Omega_{R^*})$ is related to a perturbed version $c(\sigma; y_k, \Omega_R(x_k))$. Recall that R is chosen so that x^* is the unique global minimum of f in Ω_{R^*}.

LEMMA 4.4. *Let f be twice continuously differentiable. Suppose Assumptions* 3.1 *and* 3.2 *are satisfied, and that the sequence $\{y_k\}$ is chosen as in Lemma* 4.1. *Define*

$$\mu_1(k) = \max\{\|y_k - x^*\|, \|\nabla f(y_k) - \nabla f(x^*)\|, \|\nabla^2 f(y_k) - \nabla^2 f(x^*)\|\},$$

$$\mu_2(k) = \inf\{r \,|\, \Omega_R(x_k) \subset \Omega_{R^*} + rB\}.$$

Suppose that, for some $a > 0$ and $\alpha \in [2, 3)$,

$$c(\sigma; x^*, \Omega_{R^*}) \geqq a\sigma^\alpha \quad \text{for all } \sigma > 0.$$

Then for each $\bar{a} \in (0, a)$ *and* $\bar{\sigma} > 0$ *there is a* δ *that satisfies* $0 < \delta < \bar{\sigma}$ *such that if* $\mu_1(k) + \mu_2(k) \leqq \delta$ *then*

$$c(\sigma; y_k, \Omega_R(x_k)) \geqq \bar{a}\sigma^\alpha \quad \text{for all } \sigma > \bar{\sigma}.$$

Proof. The proof follows from Theorem 3.1 of [14] and Theorem 1 of [8]. □

LEMMA 4.5. *Suppose the assumptions of Lemma* 4.4 *hold. Then for each* $\bar{a} \in (0, a)$ *and each* $\bar{\sigma} > 0$ *there is an integer* $\bar{K}(\bar{a}, \bar{\sigma})$ *such that for all* $k \geqq \bar{K}(\bar{a}, \bar{\sigma})$,

$$c(\sigma; y_k, \Omega_R(x_k)) \geqq \bar{a}\sigma^\alpha \quad \text{for all } \sigma > \bar{\sigma}.$$

Proof. By Assumption 3.1, $\mu_2(k) \to 0$ as $k \to \infty$, and by the continuity properties of f and the fact that $y_k \to x^*$, also $\mu_1(k) \to 0$ as $k \to \infty$. The result follows from Lemma 4.4 by choosing $\bar{K}(\bar{a}, \bar{\sigma})$ so that $\mu_1(k) + \mu_2(k) \leqq \delta$ for all $k \geqq \bar{K}(\bar{a}, \bar{\sigma})$. □

These results show that, outside of a small ball around y_k, the increase rate of $c(\sigma; y_k, \Omega_R(x_k))$ matches that of $c(\sigma; x^*, \Omega_{R^*})$. Finally we use this fact to find a relationship between $\|x_k - y_k\|$ and $\|x_{k+1} - y_k\|$.

THEOREM 4.6. *Suppose the assumptions of Lemma* 4.4 *hold, and that the sequence* $\{x_k\}$ *is generated by* (3.1) *with* $B_k = \nabla^2 f(x_k)$. *Then for each* $\bar{\sigma} > 0$ *there is an integer* $K(\bar{\sigma})$ *such that for* $k \geqq K(\bar{\sigma})$,

$$\|x_{k+1} - y_k\| \leqq \gamma \|x_k - y_k\|^{2/(\alpha-1)},$$

provided

$$\bar{\sigma} \leqq \|x_k - y_k\| \leqq \varepsilon, \qquad \bar{\sigma} \leqq \|x_{k+1} - y_k\| \tag{4.7}$$

where ε *and* γ *are positive constants that do not depend on* $\bar{\sigma}$.

Proof. Choose $K(\bar{\sigma}) = \bar{K}(a/2, \bar{\sigma})$. Then by Lemma 4.5, $c(\sigma; y_k, \Omega_R(x_k)) \geqq (a/2)\sigma^\alpha$ for all $\sigma > \bar{\sigma}$. The result then follows directly from Theorem 2.3 of [14]. □

The assumption (4.7) may preclude the use of Theorem 4.6 when the convergence of y_k to x^* is slow. This is because, by the conditions of Lemma 4.4,

$$\|x^* - y_k\| \leqq \mu_1(k) \leqq \delta < \bar{\sigma},$$

and so (4.7) certainly cannot apply if

$$\|x_k - y_k\| \leqq \|x^* - y_k\|, \qquad \|x_{k+1} - y_k\| \leqq \|x^* - y_k\|. \tag{4.8}$$

Finally, we give examples to show that the assumptions of Theorems 4.3 and 4.6 are independent.

Example 4.7. (i) The problem

$$\min z_2^4 \quad \text{s.t. } z_2 \geqq z_1, \quad z_2 \geqq -z_1$$

has solution $x^* = (0, 0)^T$. Assume $\Omega(x) = \Lambda(x)$, and so $\Omega(x_k) \equiv \Omega = \{y \mid y_2 \geqq y_1, y_2 \geqq -y_1\}$, and $y_k \equiv 0$. Then $\operatorname{lin}(T(y_k; \Omega(x_k))) = \{(0, 0)^T\}$ for all k. Also $(0, 0)^T$ is a quasi-polyhedral face of $\Omega(x_k)$ and so the conditions of Theorem 4.3 are trivially satisfied. Clearly $\Omega_* = \Omega$ and so from the definition of c,

$$c(\sigma; 0, \Omega_*) = \inf_{\|x\| \geqq \sigma, x \in \Omega_*} \langle \nabla f(0), x \rangle + \tfrac{1}{2}\langle x, \nabla^2 f(0) x \rangle = 0;$$

therefore the conditions of Theorem 4.6 cannot be met.

(ii) Consider

$$\min (-\cos z_2) \quad \text{s.t. } z_2 \geqq |z_1|^{4/3}.$$

The solution is again $x^* = (0, 0)^T$. Define

$$\Omega(z) = \{y \mid y_2 \geqq -z_2^2 + |y_1|^{4/3}\}.$$

Denote the components of the current iterate x_k by $x_{k,1}$ and $x_{k,2}$. The subproblem in (3.1) at x_k with $B_k = \nabla^2 f(x_k)$ is

$$\min_{x_{k+1,2}} (\sin x_{k,2})(x_{k+1,2} - x_{k,2}) + \tfrac{1}{2}(\cos x_{k,2})(x_{k+1,2} - x_{k,2})^2$$

$$\text{s.t.} \quad x_{k+1,2} \geqq -x_{k,2}^2 + |x_{k+1,1}|^{4/3}.$$

So if we set $x_{0,1} = 0$, $x_{0,2}$ small, the sequence generated by (3.1) satisfies

$$x_{k+1,1} = 0, \qquad x_{k+1,2} \approx -\tfrac{1}{3}x_{k,2}^3.$$

Also $x^* = (0, 0)^T$ is always feasible, and so we can choose $y_k \equiv (0, 0)^T$. Then $\operatorname{lin}(T(y_k; \Omega(x_k))) = \mathbb{R}^n$, but clearly (4.6) is not satisfied. However, since $\Omega(x_k) \to \Omega(0) = \Omega$, Assumption 3.1 holds. Also c is defined by

$$c(\sigma; 0, \Omega) = \inf_{\substack{\|x\| \geqq \sigma \\ x \in \Omega}} \tfrac{1}{2} z_2^2.$$

The infimum is clearly attained when $z_2 = |z_1|^{4/3}$, and $z_1^2 + z_2^2 = \sigma^2$; hence

$$z_1^2 + |z_1|^{8/3} = \sigma^2 \Rightarrow |z_1| \approx \sigma \quad \text{for small } \sigma.$$

Hence

$$c(\sigma; 0, \Omega) \approx \tfrac{1}{2}\sigma^{8/3},$$

and so the condition on c can be satisfied for some choice of $a > 0$, $\alpha \in (2, 3)$, $R > 0$. Also note that for the sequence generated above, $\|x^* - y_k\| \equiv 0$, and so no problems of form (4.8) arise. Hence the conditions of Theorem 4.6 hold.

5. Discussion. The weakness of the assumptions on the sequence of sets $\Omega(x_k)$ (Assumptions 3.1) gives rise to consideration of some alternatives to the standard choice $\Lambda(x_k)$, which is often used in algorithms for problem (1.1)-(1.3). For instance, instead of forcing a linearized equality constraint to hold exactly, as in

$$\langle z - x_k, \nabla e_i(x_k) \rangle + e_i(x_k) = 0,$$

we could instead allow some "slack," as in

$$\langle z - x_k, \nabla e_i(x_k) \rangle + e_i(x_k) \geqq -\eta_{ik}, \qquad \langle z - x_k, \nabla e_i(x_k) \rangle + e_i(x_k) \leqq \eta_{ik}$$

where $\eta_{ik} \geqq 0$ and $\eta_{ik} \to 0$ as $k \to \infty$. If $\eta_{ik} = \tau_k |e_i(x_k)|$ with $\tau_k \in (0, 1)$, then (5.1) is enforcing a "linearized reduction" in e_i. Similarly, we may allow some "slack" in the linearized inequalities, as in

$$\langle z - x_k, \nabla c_j(x_k) \rangle + c_j(x_k) \geqq -\bar{\eta}_{jk}$$

where $\bar{\eta}_{jk} \geqq 0$ with $\bar{\eta}_{jk} \to 0$ as $k \to \infty$.

The use of an active set strategy, in which $\Omega(x_k)$ is defined as

$$\Omega(x_k) = \left\{ z \,\middle|\, \begin{array}{ll} \langle z - x_k, \nabla c_j(x_k) \rangle + c_j(x_k) = 0, & j \in \mathcal{A}_k \\ \langle z - x_k, \nabla e_i(x_k) \rangle + e_i(x_k) = 0, & i = 1, \cdots, m_E \end{array} \right.,$$

also satisfies Assumption 3.1, provided that the sequence of active sets $\mathcal{A}_k$ only changes finitely often. In a similar vein we could use the definition

$$\Omega(x_k) = \left\{ z \,\middle|\, \begin{array}{ll} \langle z - x_k, \nabla c_j(x_k) \rangle + c_j(x_k) \geqq 0, & j \in \bar{\mathcal{A}}_k \\ \langle z - x_k, \nabla e_i(x_k) \rangle + e_i(x_k) = 0, & i = 1, \cdots, m_E \end{array} \right.$$

where $\bar{\mathcal{A}}_k \subset \{1, \cdots, m_I\}$ excludes those constraints that obviously will not be active at x^* (thereby saving the cost of evaluating these constraints and their gradients).

The case in which $\nabla c_j(x_k)$ and $\nabla e_i(x_k)$ are *approximated* rather than calculated exactly is also covered, provided that the approximations are asymptotically exact as $k \to \infty$. For example, to avoid recalculation of ∇e_i and ∇c_j at every step, the same values might be used for more than one iteration. Finite-difference approximations might also be considered where appropriate.

A final possibility is the use of a trust region bound on the step size at each iteration, as in

$$\Omega(x_k) = \Lambda(x_k) \cap (x_k + R_k B)$$

where $R_k > 0$. It is clear that $\Omega(x_k)$ may be empty if R_k is too small and $x_k \notin \Omega(x_k)$. Hence in Vardi [16] this is combined with the use of slackness in the equality constraints of the form

$$\langle z - x_k, \nabla e_i(x_k) \rangle + \alpha e_i(x_k) = 0$$

where $\alpha \in (0, 1]$, to produce a globally convergent SQP method. This overall strategy is also covered by our earlier analysis.

Acknowledgment. I wish to acknowledge the detailed comments of a referee which greatly improved the paper.

REFERENCES

[1] H. ATTOUCH, *Variational convergence of functions and operators*, in Research Notes in Mathematics, Pitman, London, 1985.

[2] J. V. BURKE AND J. J. MORÉ, *On the identification of active constraints*, SIAM J. Numer. Anal., 25 (1988), pp. 1197-1211.

[3] P. H. CALAMAI AND J. J. MORÉ, *Projected gradient methods for linearly constrained problems*, Math. Programming, 39 (1987), pp. 93-116.

[4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[5] J. C. DUNN, *Newton's method and the Goldstein step-length rule for constrained minimization problems*, SIAM J. Control Optim., 18 (1980), pp. 659-674.

[6] ———, *Global and asymptotic convergence rate estimates for a class of projected gradient processes*, SIAM J. Control Optim., 19 (1981), pp. 368-400.

[7] ———, *On the convergence of projected gradient processes to singular critical points*, J. Optim. Theory Appl., 55 (1987), pp. 203-216.

[8] J. C. DUNN AND E. SACHS, *The effect of perturbations on the convergence rates of optimization algorithms*, Appl. Math. Optim., 10 (1983), pp. 143-157.

[9] R. FLETCHER, *Practical Methods of Optimization, Vol. 2: Constrained Optimization*, John Wiley, New York, 1981.

[10] F. J. GOULD AND J. W. TOLLE, *A necessary and sufficient qualification for constrained optimization*, SIAM J. Appl. Math., 20 (1971), pp. 164-172.

[11] S. M. ROBINSON, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130-143.

[12] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[13] ———, *Lipschitzian properties of multifunctions*, Nonlinear Anal., 9 (1985), pp. 867-885.

[14] E. SACHS, *Newton's method for singular constrained optimization*, Appl. Math. Optim., 11 (1984), pp. 247-276.

[15] G. SALINETTI AND R. J.-B. WETS, *On the convergence of sequences of convex sets in finite dimensions*, SIAM Rev., 21 (1979), pp. 18-33.

[16] A. VARDI, *A trust region algorithm for equality constrained minimization: convergence properties and implementation*, SIAM J. Numer. Anal., 22 (1985), pp. 575-591.

[17] R. B. WILSON, *A simplicial algorithm for concave programming*, Ph.D. thesis, Graduate School of Business Administration, Harvard University, Cambridge, MA, 1963.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 27-42, January 1989

003

THE REGULAR FREE-ENDPOINT LINEAR QUADRATIC PROBLEM WITH INDEFINITE COST*

HARRY L. TRENTELMAN†

Abstract. This paper studies an open problem in the context of linear quadratic optimal control, the free-endpoint regular linear quadratic problem with *indefinite* cost functional. It is shown that the optimal cost for this problem is given by a particular solution of the algebraic Riccati equation. This solution is characterized in terms of the geometry on the lattice of all real symmetric solutions of the algebraic Riccati equation as developed by Willems [*IEEE Trans. Automat. Control*, 16 (1971), pp. 621-634] and Coppel [*Bull. Austral. Math. Soc.*, 10 (1974), pp. 377-401]. A necessary and sufficient condition is established for the existence of optimal controls. This condition is stated in terms of a subspace inclusion involving the extremal solutions of the algebraic Riccati equation. The optimal controls are shown to be generated by a feedback control law. Finally, the results obtained are compared with "classical" results on the linear quadratic regulator problem.

Key words. linear quadratic optimal control, indefinite cost functional, free-endpoint problem

AMS(MOS) subject classifications. 93C05, 93C35, 93C60

1. Introduction. In this paper we are concerned with regular, infinite-horizon linear quadratic optimal control problems in which the cost functional is the integral of an *indefinite* quadratic form.

In most of the existing literature on the regular linear quadratic (LQ) problem, it is explicitly assumed that the quadratic form in the cost functional, apart from being positive definite in the control variable alone, is positive semidefinite in the control and state variables simultaneously. In fact, under this semidefiniteness assumption the LQ problem has become quite standard and is treated in many basic textbooks in the field of systems and control [1], [2], [9], [21]. Often a distinction is made between two versions of the problem, the *fixed-endpoint* version and the *free-endpoint* version. In the fixed-endpoint version it is necessary to minimize the cost functional under the constraint that the optimal state trajectory should converge to zero as time tends to infinity, while in the free-endpoint version it is only necessary to minimize the cost functional. For the case that the quadratic form in the cost functional is positive semidefinite both versions of the regular LQ problem are well-understood and completely satisfactory solutions of these problems are available.

Surprisingly, however, for the most general formulation of the regular LQ problem, that is, the case that the quadratic form in the cost functional is indefinite, a satisfactory treatment does not yet exist. In this case we can again distinguish between the fixed-endpoint version and the free-endpoint version. While for the fixed-endpoint version a complete solution has been described in [17] (see also [14]), the free-endpoint version has only been considered in [17] under a very restrictive assumption. Thus we see that, up to now, the free-endpoint regular LQ problem with indefinite cost functional has been an open problem. In the present paper we shall fill up this gap and present a fairly complete solution to this problem.

It is well known [12], [19] that for the free-endpoint regular LQ problem with positive semidefinite cost functional, the optimal cost is given by the smallest positive semidefinite real symmetric solution of the algebraic Riccati equation. We will see that this statement is no longer valid in general if the cost functional is the integral of an

* Received by the editors November 2, 1987; accepted for publication (in revised form) March 28, 1988.
† Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, the Netherlands.

indefinite quadratic form. It will be shown, however, that in this case also the optimal cost is given by a solution of the algebraic Riccati equation. This particular solution will be characterized in terms of the geometry on the set of all real symmetric solutions of the algebraic Riccati equation as described in [17] and [4].

Another well-known fact is that, for the free-endpoint regular LQ problem with positive semidefinite cost functional, the *existence* of optimal controls is never an issue: under the assumption that the underlying system is controllable, for this problem unique optimal controls always exist for all initial conditions. This is in contrast with the fixed-endpoint LQ problem, where the existence of optimal controls for all initial conditions depends on the "gap" of the algebraic Riccati equation (i.e., the difference between the largest and smallest solutions of the Riccati equation). In this paper we will see that also, for the free-endpoint regular LQ problem with *indefinite* cost functional, optimal controls no longer need to exist for all initial conditions! Moreover, we will establish a necessary and sufficient condition in terms of the "gap" of the algebraic Riccati equation for the existence of optimal controls for all initial conditions. We will show that for the particular case that the cost functional is positive semidefinite this condition is always satisfied, thus explaining the fact that in this special case optimal controls always exist. Finally, we will show that also in the indefinite case the optimal controls for the free-endpoint regular LQ problem, if they exist, are given by a feedback control law.

The outline of this paper is as follows. In the remainder of this section we will introduce most of the notational conventions that will be used. In § 2 we give formulations of both the free-endpoint and fixed-endpoint regular LQ problems that we shall be dealing with. In § 3 we will briefly recall the most important facts that we need on the geometry of the set of all real symmetric solutions to the algebraic Riccati equation as developed in [17] and [4]. In § 4 we will state the solution to the fixed endpoint regular LQ problem with indefinite cost as established in [17]. Also, we will state its (incomplete) counterpart, the solution to the free-endpoint regular LQ problem with positive semidefinite cost functional. Then in § 5 we will state and prove our main theorem, a solution to the free-endpoint regular LQ problem. In order to establish a proof of this theorem we will state and prove a series of smaller lemmas. In § 6 we will show how the "classical" results on the free-endpoint regular LQ problem with positive semidefinite cost functional can be reobtained as a special case of our general solution. We will close this paper in § 7 with some concluding remarks.

We use the following notational conventions. For a given $n \times n$ matrix A its set of eigenvalues will be denoted by $\sigma(A)$. If V is a subspace of $\mathbb{R}^n$ and A is an $n \times n$ matrix then $A|V$ will denote the restriction of A to V. V will be called A-invariant if $AV \subset V$. In this case $\sigma(A|V)$ will denote the set of eigenvalues of $A|V$ and $\sigma(A|\mathbb{R}^n/V)$ will denote the set of eigenvalues of the mapping induced by A in the factor space $\mathbb{R}^n/V$ (see [21]). We will denote subsets of $\mathbb{C}$ by $\mathbb{C}^- := \{s \in \mathbb{C} | \operatorname{Re} s = 0\}$, $\mathbb{C}^0 := \{s \in \mathbb{C} | \operatorname{Re} s = 0\}$, and $\mathbb{C}^+ := \{s \in \mathbb{C} | \operatorname{Re} s > 0\}$. Given a real monic polynomial p there is a unique factorization $p = p_- \cdot p_0 \cdot p_+$ into real monic polynomials with p_-, p_0, and p_+ having all roots in $\mathbb{C}^-$, $\mathbb{C}^0$, and $\mathbb{C}^+$, respectively. If A is a real $n \times n$ matrix and if p denotes its characteristic polynomial then we define $X^-(A) := \ker p_-(A)$, $X^0(A) := \ker p_0(A)$, and $X^+(A) := \ker p_+(A)$. These subspaces are A-invariant and the restriction of A to $X^-(A)(X^0(A), X^+(A))$ has characteristic polynomial $p_-(p_0, p_+)$.

A subset $\mathbb{C}_g$ of $\mathbb{C}$ will be called symmetric if $a + bi \in \mathbb{C}_g \Leftrightarrow a - bi \in \mathbb{C}_g$. If $\mathbb{C}_g$ is given then we define $\mathbb{C}_b := \mathbb{C} \backslash \mathbb{C}_g$. If A is a real $n \times n$ matrix and if p is its characteristic polynomial then, again, p can be factored uniquely into $p = p_g \cdot p_b$, where p_g and p_b are real monic polynomials with all roots in $\mathbb{C}_g$ and $\mathbb{C}_b$, respectively. We denote

$X_g(A) := \ker p_g(A)$ and $X_b(A) := \ker p_b(A)$. Again these subspaces are A-invariant and the restriction of A to $X_g(A)(X_b(A))$ has characteristic polynomial $p_g(p_b)$. In fact, the subspace $X_g(A)(X_b(A))$ is equal to the linear span of all generalized eigenvectors of A corresponding to its eigenvalues in $\mathbb{C}_g(\mathbb{C}_b)$. Alternatively, $X_g(A)(X_b(A))$ is equal to the largest A-invariant subspace V of $\mathbb{R}^n$ such that $\sigma(A|V) \subset \mathbb{C}_g(\mathbb{C}_b)$.

If, in addition to A, a real $p \times n$ matrix C is given, then we denote

$$\langle \ker C|A\rangle := \bigcap_{i=1}^{n} \ker CA^{i-1},$$

the unobservable subspace of (C, A) [21, § 3.2]. Given a symmetric subset $\mathbb{C}_g$ of $\mathbb{C}$ we denote

$$X_{\text{det}} := \langle \ker C|A\rangle \cap X_b(A),$$

the undetectable subspace of (C, A) with respect to $\mathbb{C}_g$. The pair (C, A) is called detectable with respect to $\mathbb{C}_g$ if A is $(\mathbb{C}_{g-})$ stable on the unobservable subspace of (C, A), i.e., if

$$\langle \ker C|A\rangle \subset X_g(A)$$

(see [21, § 3.6]). It is easy to see that (C, A) is detectable if and only if $X_{\text{det}} = 0$. Also, (C, A) is detectable if and only if for all $\lambda \in \mathbb{C}_b$ we have $\ker (A - \lambda I) \cap \ker C = 0$ (see [15]).

In order to be rigorous on the interpretation of the cost functionals that will be considered in this paper, we will now explain what we mean by the statement that the limit of a function *exists in* $\mathbb{R}^e$. Let $\mathbb{R}^e := \mathbb{R} \cup \{-\infty, +\infty\}$. Given $f: \mathbb{R} \to \mathbb{R}$ we say that $\lim_{t\to\infty} f(t)$ exists if it is equal to a real number in the usual sense. We say that $\lim_{t\to\infty} f(t) = -\infty(+\infty)$ if for all $r \in \mathbb{R}$ there exists $T \in \mathbb{R}$ such that $t \geqq T$ implies $f(t) \leqq r(\geqq r)$. Then we say that $\lim_{t\to\infty} f(t)$ exists in $\mathbb{R}^e$ if it exists, is equal to $-\infty$, or is equal to $+\infty$.

If M is a real $n \times n$ matrix and V is a subspace of $\mathbb{R}^n$, then we define $M^{-1}V := \{x \in \mathbb{R}^n | Mx \in V\}$. If V is a subspace of $\mathbb{R}^n$ then $V^\perp$ denotes its orthogonal complement with respect to the standard Euclidean inner product.

Finally, we will denote by $L_{2,\text{loc}}(\mathbb{R}^+)$ the space of all measurable vector-valued functions on $\mathbb{R}^+$ that are square integrable over all finite intervals in $\mathbb{R}^+$. $L_2(\mathbb{R}^+)$ denotes the space of all measurable vector-valued functions on $\mathbb{R}^+$ that are square integrable over $\mathbb{R}^+$. Finally, $L_\infty(\mathbb{R}^+)$ denotes the space of all measurable vector-valued functions on $\mathbb{R}^+$ that are essentially bounded on $\mathbb{R}^+$. Here, $\mathbb{R}^+ := \{t \in \mathbb{R} | t \geqq 0\}$.

2. The regular LQ-problem. Consider the finite-dimensional linear time-invariant system

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0. \tag{2.1}$$

Here, x and u are assumed to take their values in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. A and B are real $n \times n$ and $n \times m$ matrices, respectively. It will be a standing assumption that (A, B) is controllable. We shall consider optimization problems of the type

$$\inf \int_0^\infty \omega(x, u)\, dt, \tag{2.2}$$

where $\omega(x, u)$ is a real quadratic form on $\mathbb{R}^n \times \mathbb{R}^m$ defined by $\omega(x, u) := u^T R u + 2u^T S x + x^T Q x$. Here R, S, and Q are assumed to be real matrices such that $R = R^T$ and $Q = Q^T$. As in [17], no a priori definiteness conditions are imposed on

the form ω. For a given control function $u \in L_{2,\mathrm{loc}}(\mathbb{R}^+)$, let $x(x_0, u)$ denote the state trajectory of (2.1) and if $T \geqq 0$ let

$$J_T(x_0, u) := \int_0^T \omega(x(x_0, u)(t), u(t))\, dt.$$

We now explain how (2.2) should be interpreted. First, we specify two classes of control functions with respect to which the infimization in (2.2) should be performed. Define

$$U(x_0) := \{u \in L_{2,\mathrm{loc}}(\mathbb{R}^+) \mid \lim_{T\to\infty} J_T(x_0, u) \text{ exists in } \mathbb{R}^e\},$$

$$U_s(x_0) := \{u \in U(x_0) \mid \lim_{t\to\infty} x(x_0, u)(t) = 0\}.$$

Note that, due to the assumption that (A, B) is controllable, we have $U(x_0) \neq \varnothing$ and $U_s(x_0) \neq \varnothing$ for all $x_0 \in \mathbb{R}^n$. For $u \in U(x_0)$ we define

$$J(x_0, u) := \lim_{T\to\infty} J_T(x_0, u). \tag{2.3}$$

We note that $J(x_0, u) \in \mathbb{R}^e$. Now, define

$$V_f^+(x_0) := \inf\{J(x_0, u) \mid u \in U(x_0)\}, \tag{2.4a}$$

$$V^+(x_0) := \inf\{J(x_0, u) \mid u \in U_s(x_0)\}, \tag{2.4b}$$

the optimal cost for the free-endpoint problem and fixed-endpoint problem, respectively. By the fact that (A, B) is controllable we have that $V_f^+(x_0)$, $V^+(x_0) \in \mathbb{R} \cup \{-\infty\}$ for all $x_0 \in \mathbb{R}^n$. Following [17], we want to exclude the situation that for certain initial conditions x_0 the values (2.4a) or (2.4b) become equal to $-\infty$. It can be shown that a necessary condition for $V_f^+(x_0) > -\infty$ and $V^+(x_0) > -\infty$ for all x_0 to hold is that $R \geqq 0$ (see [17], [12]). In this paper a standing assumption will be that $R > 0$. Under this assumption the LQ problems defined by (2.4) are called *regular.*

The fixed-endpoint regular LQ problem, defined by (2.4b), was completely resolved in [17] (see also [14]). There, a satisfactory characterization was given for the optimal cost, necessary and sufficient conditions were given for the existence of optimal controls for all initial conditions, and these optimal controls were given in the form of a state-feedback control law. The problems of how to calculate the optimal cost for the free-endpoint regular LQ problem (2.4a), to state necessary and sufficient conditions for the existence of optimal controls, and to calculate these optimal controls have up to now been open. In this paper we will consider these problems.

3. Geometry of the algebraic Riccati equation. A central role in infinite horizon regular linear quadratic control problems is played by the algebraic Riccati equation (ARE)

$$A^TK + KA + Q - (KB + S^T)R^{-1}(B^TK + S) = 0. \tag{3.1}$$

Let Γ denote the set of all real symmetric solutions of the ARE. It was shown in [17] that if Γ is nonempty then it contains a unique element K^+ and a unique element K^- such that

$$\sigma(A - BR^{-1}(B^TK^+ + S)) \subset \mathbb{C}^- \cup \mathbb{C}^0,$$

$$\sigma(A - BR^{-1}(B^TK^- + S)) \subset \mathbb{C}^+ \cup \mathbb{C}^0.$$

Moreover, K^+ and K^- have the additional property that they are the *extremal solutions* of the ARE in the sense that if $K \in \Gamma$ then $K^- \leqq K \leqq K^+$.

Let $\Delta := K^+ - K^-$. Denote $A - BR^{-1}(B^TK^+ + S)$ and $A - BR^{-1}(B^TK^- + S)$ by A^+ and A^-, respectively. If $K \in \Gamma$ define $A_K := A - BR^{-1}(B^TK + S)$. Note that $X^+(A^+) = 0$ and $X^-(A^-) = 0$. Let Ω denote the set of all A^--invariant subspaces contained in $X^+(A^-)$. The following basic theorem is a generalization by Coppel [4] of a theorem that was originally proven by Willems in [17] (see also [16], [10]).

THEOREM 3.1. *Let (A, B) be controllable, and assume that Γ is nonempty. If V is an A^--invariant subspace of $X^+(A^-)$ (that is, if $V \in \Omega$) then $\mathbb{R}^n = V \oplus \Delta^{-1}V^\perp$. There exists a bijection $\gamma : \Omega \to \Gamma$ defined by*

$$\gamma(V) := K^-P_V + K^+(I - P_V),$$

where P_V is the projector onto V along $\Delta^{-1}V^\perp$. If $K = \gamma(V)$ then

$$X^+(A_K) = V,$$

$$X^0(A_K) = \ker \Delta,$$

$$X^-(A_K) = X^-(A^+) \cap \Delta^{-1}V^\perp.$$

Among other things, the result above states that there exists a one-to-one correspondence between the set of all real symmetric solutions of the ARE and the set of all A^--invariant subspaces of $X^+(A^-)$. Following [3], if $K = \gamma(V)$ then we will say that *the solution K is supported by the subspace V.* The next theorem from [4] states that this one-to-one correspondence in fact respects the partial orderings on the sets Γ and Ω.

THEOREM 3.2. *Let (A, B) be controllable and assume that Γ is nonempty. Let K_1 and K_2 be solutions to the* ARE *supported by V_1 and V_2, respectively. Then $K_1 \leqq K_2$ if and only if $V_2 \subset V_1$.*

From the above it follows, for example, that K^- is supported by $X^+(A^-)$ and that K^+ is supported by 0.

4. Classical results. In the present section we briefly summarize the solution of the fixed-endpoint regular LQ problem with indefinite cost functional as outlined in [17]. Subsequently, we will state the well-known result on the free-endpoint regular LQ problem with *positive semidefinite* cost functional. Finally, we will discuss some of the difficulties that can be expected in trying to generalize the latter result to the case that the semidefiniteness assumption is dropped.

Consider the infimization of (2.3) over the class of inputs $U_s(x_0)$. For a given x_0 an input u^* is called *optimal* if $u^* \in U_s(x_0)$ and $J(x_0, u^*) = V^+(x_0)$. The following was proven in [17].

THEOREM 4.1. *Let (A, B) be controllable and assume that $R > 0$. Then we have the following:*

(i) *$V^+(x_0)$ is finite for all $x_0 \in \mathbb{R}^n$ if and only if the* ARE *has a real symmetric solution (i.e., $\Gamma \neq \varnothing$).*

(ii) *If $\Gamma \neq \varnothing$ then for all $x_0 \in \mathbb{R}^n$, $V^+(x_0) = x_0^TK^+x_0$.*

(iii) *If $\Gamma \neq \varnothing$ then for all $x_0 \in \mathbb{R}^n$ there exists an optimal input u^* if and only if $\Delta > 0$.*

(iv) *If $\Gamma \neq \varnothing$ and $\Delta > 0$ then for each $x_0 \in \mathbb{R}^n$ there is exactly one optimal input u^* and, moreover, this input u^* is given by the feedback control law $u^* = -R^{-1}(B^TK^+ + S)x$.*

As already mentioned, an analogue of the latter theorem for the free-endpoint case, up to now, has only been available for the case that the quadratic form $\omega(x, u)$ is positive semidefinite, i.e., for the case that $\omega(x, u) \geqq 0$ for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$. In the sequel, let $\Gamma_+ := \{K \in \Gamma \mid K \geqq 0\}$. It is well known [8], [12] that if $\omega \geqq 0$ and if (A, B) is

controllable, then the ARE has a smallest positive semidefinite real symmetric solution. More precisely, there exists a (unique) $\tilde{K}$ such that

$$\tilde{K} \in \Gamma_+, \tag{4.1}$$

$$K \in \Gamma_+ \Rightarrow \tilde{K} \leqq K. \tag{4.2}$$

The solution $\tilde{K}$ characterized by (4.1) and (4.2) plays the central role in the solution of the free-endpoint regular LQ problem with positive semidefinite cost. In the following, for a given $x_0 \in \mathbb{R}^n$ an input u^* is called *optimal* if $u^* \in U(x_0)$ and $J(x_0, u^*) = V_f^+(x_0)$.

THEOREM 4.2. *Assume that (A, B) is controllable, that $R > 0$, and that $\omega(x, u) \geqq 0$ for all $(x, u) \in \mathbb{R}^n \times \mathbb{R}^m$. Then we have the following*:

(i) *For all $x_0 \in \mathbb{R}^n$, $V_f^+(x_0) = x_0^T \tilde{K} x_0$.*

(ii) *For each $x_0 \in \mathbb{R}^n$, there is exactly one optimal input u^*, and moreover, this input u^* is given by the feedback control law $u^* = -R^{-1}(B^T\tilde{K} + S)x$.*

Proof. This follows, for example, by combining [12, Thm. 8] and the results from [1, p. 36] (see also [19]). □

We note that in this theorem the *existence* of optimal controls is no issue. In contrast with the fixed-endpoint problem, the positive semidefiniteness assumption assures that in the free-endpoint problem for every initial condition there exists an optimal control.

In trying to generalize the latter theorem to the case that ω is an arbitrary indefinite quadratic form in (x, u) (with of course, as usual, $R > 0$), the following aspects should be considered. First, due to the indefiniteness of ω, the optimal cost $V_f^+(x_0)$ no longer needs to be finite. In this paper we want to restrict ourselves to the case that $V_f^+(x_0)$ is finite for all x_0. In order to establish a condition assuring this, we state the following well-known result. For $\nu \in \mathbb{R}^m$, denote $\|\nu\|_R^2 := \nu^T R \nu$.

LEMMA 4.3. *Let $K \in \Gamma$. Then for all $u \in L_{2,\mathrm{loc}}(\mathbb{R}^+)$ and for all $T \geqq 0$, we have*

$$J_T(x_0, u) = \int_0^T \|u(t) + R^{-1}(B^T K + S)x(t)\|_R^2 \, dt + x_0^T K x_0 - x^T(T) K x(T).$$

Here, we have denoted $x(t) := x(x_0, u)(t)$.

Proof. For a proof, refer to [2] or [17]. □

In the sequel, let $\Gamma_- := \{K \in \Gamma \mid K \leqq 0\}$. From the previous lemma the following is immediate.

LEMMA 4.4. *Let (A, B) be controllable and $R > 0$. If $\Gamma_- \neq \emptyset$ then $V_f^+(x_0)$ is finite for all $x_0 \in \mathbb{R}^n$.*

Proof. $\Gamma_- \neq \emptyset$ implies that $K^- \leqq 0$. Applying the previous lemma to K^- yields $J_T(x_0, u) \geqq x_0^T K^- x_0$ for all u and $T \geqq 0$. □

Remark 4.5. In [17] it is suggested that the converse of the above lemma also holds, i.e., that finiteness of $V_f^+(x_0)$ for all x_0 implies that $\Gamma_- \neq \emptyset$. We were able neither to establish a proof nor to construct a counterexample to this assertion. We were, however, able to relate the condition $\Gamma_- \neq \emptyset$ to an equivalent one in terms of the quantities $J_T(x_0, u)$ in a slightly different way. Indeed, if (A, B) is controllable and $R > 0$ then the following equivalence can be proven:

$$\Gamma_- \neq \emptyset \Leftrightarrow \inf\{\liminf_{T \to \infty} J_T(x_0, u) \mid u \in L_{2,\mathrm{loc}}(\mathbb{R}^+)\} \text{ is finite for all } x_0 \in \mathbb{R}^n. \tag{4.3}$$

Note that if we could prove the above equivalence with $L_{2,\mathrm{loc}}(\mathbb{R}^+)$ replaced by $U(x_0)$ we would be done. Indeed, for $u \in U(x_0)$ we have $\liminf_{T\to\infty} J_T(x_0, u) = \lim_{T\to\infty} J_T(x_0, u) = J(x_0, u)$, so the infimum in (4.3) would then be equal to $V_f^+(x_0)$. We close this remark by concluding that finding tractable necessary and sufficient

conditions for the finiteness of V_f^+ remains a difficult open problem (see also [18], [11], and [13]).

A final point we want to make here is that for the free-endpoint problem with indefinite cost, even if the optimal cost is finite for all initial conditions, it is not true in general that optimal controls *exist* for all initial conditions. We will illustrate this in the example below. It should therefore be clear that part of our problem is to formulate necessary and sufficient conditions for the existence of these optimal controls (as was also done in Theorem 4.1(iii)).

Example 4.6. Consider the controllable system $\dot{x} = -x + u$, $x(0) = x_0$ with indefinite cost functional

$$J(x_0, u) = \int_0^\infty -x(t)^2 + u(t)^2 \, dt,$$

that is, take $A = -1$, $B = 1$, $Q = -1$, $S = 0$, and $R = 1$. The corresponding ARE is given by $-2K - K^2 - 1 = 0$. Consequently, $K^- = K^+ = -1$. We claim that $V_f^+(x_0) = -x_0^2$. We will show this "from first principles." Let $u \in L_{2,\mathrm{loc}}(\mathbb{R}^+)$. For every $T \geqq 0$ we have

$$\begin{aligned}\int_0^T -x^2 + u^2 \, dt &= \int_0^T (x-u)^2 \, dt + 2\int_0^T x(-x+u) \, dt \\ &= \int_0^T (x-u)^2 \, dt + 2\int_0^T x\dot{x} \, dt = \int_0^T (x-u)^2 \, dt + x^2(T) - x_0^2.\end{aligned}$$

Consequently, $J(x_0, u) \geqq -x_0^2$ for all $u \in U(x_0)$. On the other hand, for $\varepsilon > 0$ define $u = (1-\varepsilon)x$. Then $\dot{x} = -\varepsilon x$ and

$$J(x_0, u) = [(1-\varepsilon)^2 - 1]x_0^2 \int_0^\infty e^{-2\varepsilon t} \, dt = -x_0^2 + \frac{\varepsilon}{2} x_0^2.$$

It follows that $V_f^+(x_0) = \inf\{J(x_0, u) | u \in U(x_0)\} = -x_0^2$. Thus, we see that the optimal cost is finite (as could also be deduced from the fact that $K^- = -1 \leqq 0$). We claim, however, that *no optimal control exists*! Indeed, assume u^* is optimal. Let x^* be the corresponding trajectory. We have

$$-x_0^2 = J(x_0, u^*) = -x_0^2 + \lim_{T\to\infty} \left(\int_0^T (x^* - u^*)^2 \, dt + x^*(T)^2 \right).$$

From this it follows that $\int_0^\infty (x^* - u^*)^2 \, dt = 0$ and that, consequently, $u^* = x^*$. However, using this feedback control law yields $J(x_0, u^*) = 0$. If $x_0 \neq 0$ this yields a contradiction.

5. The free-endpoint regular LQ-problem with indefinite cost. In this section we will resolve the free-endpoint version of the regular LQ problem with indefinite cost functional. In the sequel, an important role will be played by the subspace

$$N := \langle \ker K^- | A^- \rangle \cap X^+(A^-). \tag{5.1}$$

By definition of A^- it is immediately clear that, in fact,

$$N = \langle \ker K^- | A - BR^{-1}S \rangle \cap X^+(A - BR^{-1}S). \tag{5.2}$$

Obviously, N is equal to the undetectable subspace of (K^-, A^-) with respect to the stability set $\mathbb{C}_g = \mathbb{C}^- \cup \mathbb{C}^0$. We also note that N is an A^--invariant subspace of $X^+(A^-)$. By Theorem 3.1, N corresponds to a real symmetric solution $\gamma(N)$ of the ARE. Let P_N be the projector onto N along $\Delta^{-1}N^\perp$. Then this solution $\gamma(N)$ is given by

$$K_f^+ := \gamma(N) = K^- P_N + K^+(I - P_N). \tag{5.3}$$

It will turn out that K_f^+, the solution of the ARE supported by the subspace N, is the bottleneck in the problem we want to resolve. We will show that the optimal cost for the free-endpoint problem is obtained from K_f^+ and that the optimal controls, if they exist, are given by the feedback control law $u = -R^{-1}(B^T K_f^+ + S)x$. Before stating the exact result we first give an intuitive argument as to exactly why the subspace N given by (5.1) is the "right" supporting subspace. The argument is as follows. First recall that if $\omega \geqq 0$, then the optimal cost for the free-endpoint problem is obtained from the smallest positive semidefinite solution of the ARE (see Theorem 4.2). Now, it can be shown that, again if $\omega \geqq 0$, $K = \gamma(V)$ is positive semidefinite if and only if $V \subset \ker K^-$ (see Theorem 6.2). Consequently, if $\omega \geqq 0$ then the optimal cost is obtained from the smallest solution $K = \gamma(V)$ of the ARE such that $V \subset \ker K^-$. Now, our choice to consider exactly the subspace N given by (5.1) is based on the guess that the latter statement is also valid if ω is indefinite. Note that K_f^+ is indeed the smallest solution of ARE for which its supporting subspace is contained in $\ker K^-$: if $K = \gamma(V)$ is such that $V \subset \ker K^-$ then, since V is A^--invariant, we must have $V \subset \langle \ker K^- | A^- \rangle$ (the latter being the largest A^--invariant subspace in $\ker K^-$). Also, $V \subset X^+(A^-)$. Thus, $V \subset N$. Then it follows from Theorem 3.2 that $K_f^+ \leqq K$. The following theorem is the main result of this paper.

THEOREM 5.1. *Let (A, B) be controllable and assume that $R > 0$. Then we have the following:*

(i) *$V_f^+(x_0)$ is finite for all $x_0 \in \mathbb{R}^n$ if the* ARE *has a negative semidefinite real symmetric solution (i.e., $\Gamma_- \neq \varnothing$).*

(ii) *If $\Gamma_- \neq \varnothing$ then for all $x_0 \in \mathbb{R}^n$, $V_f^+(x_0) = x_0^T K_f^+ x_0$.*

(iii) *If $\Gamma_- \neq \varnothing$ then for all $x_0 \in \mathbb{R}^n$ there exists an optimal input u^* if and only if $\ker \Delta \subset \ker K^-$.*

(iv) *If $\Gamma_- \neq \varnothing$ and if $\ker \Delta \subset \ker K^-$, then for each $x_0 \in \mathbb{R}^n$ there is exactly one optimal input u^* and, moreover, this input is given by the feedback control law $u^* = -R^{-1}(B^T K_f^+ + S)x$.*

In the remainder of this section we will establish a proof of this theorem. In order to streamline this proof, we will state some of the most important ingredients as separate lemmas. In the first two lemmas, we will formulate some general structural properties of linear systems.

LEMMA 5.2. *Consider the system $\dot{x} = Ax + v$, $y = Cx$. Assume that (C, A) is observable. Let $v \in L_2(\mathbb{R}^+)$, $y \in L_\infty(\mathbb{R}^+)$. Then for every initial condition x_0 we have $x \in L_\infty(\mathbb{R}^+)$.*

Proof. Since (C, A) is observable there exists a matrix L such that $\sigma(A + LC) \subset \mathbb{C}^-$. Obviously, x satisfies the differential equation

$$\dot{x} = (A + LC)x - Ly + v, \qquad x(0) = x_0.$$

Using the variations of constants formula, together with some straightforward estimates, it is then easily verified that $x \in L_\infty(\mathbb{R}^+)$. □

Using the previous lemma we arrive at the following result that will be one of the main instruments in the proof of Theorem 5.1.

LEMMA 5.3. *Consider the system $\dot{x} = Ax + v$, $y = Cx$. Let $\mathbb{C}_g$ be a symmetric subset of $\mathbb{C}$. Assume that (C, A) is detectable with respect to $\mathbb{C}_g$. Let the state space $\mathbb{R}^n$ be decomposed into $\mathbb{R}^n = X_1 \oplus X_2$, where X_1 is A-invariant. In this decomposition, let $x = \binom{x_1}{x_2}$. Assume that $\sigma(A|X_1) \subset \mathbb{C}_g$ and $\sigma(A|\mathbb{R}^n/X_1) \subset \mathbb{C}_b$. Then for every initial condition x_0 we have: if $v \in L_2(\mathbb{R}^+)$ and $y \in L_\infty(\mathbb{R}^+)$ then $x_2 \in L_\infty(\mathbb{R}^+)$.*

Proof. We claim that, in fact, $X_1 = X_g(A)$. Indeed, the fact that $X_1 \subset X_g(A)$ is immediate. Denote $\sigma_0 := \sigma(A|X_g(A)/X_1)$. Then $\sigma_0 \subset \sigma(A|X_g(A)) \subset \mathbb{C}_g$. Also, $\sigma_0 \subset \sigma(A|\mathbb{R}^n/X_1) \subset \mathbb{C}_b$. This can only be the case if $\sigma_0 = \varnothing$ or, equivalently, if $X_1 = X_g(A)$.

By the fact that (C, A) is detectable with respect to $\mathbb{C}_g$ we may therefore conclude that $\langle \ker C|A\rangle \subset X_1$. Decompose $X_1 = X_{11} \oplus X_{12}$, with $X_{11} := \langle \ker C|A\rangle$ and X_{12} arbitrarily. Accordingly, let $x_1 = \binom{x_{11}}{x_{12}}$. We then have $\mathbb{R}^n = X_{11} \oplus X_{12} \oplus X_2$ with $x = (x_{11}^T, x_{12}^T, x_2^T)^T$. In this decomposition, let

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}, \quad C = (0, C_2, C_3), \quad \nu = \begin{pmatrix} \nu_{11} \\ \nu_{12} \\ \nu_2 \end{pmatrix}.$$

Obviously, the system

$$\left((C_2, C_3), \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \right)$$

is observable. Moreover,

$$\begin{pmatrix} \dot{x}_{12} \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{pmatrix} \begin{pmatrix} x_{12} \\ x_2 \end{pmatrix} + \begin{pmatrix} \nu_{12} \\ \nu_2 \end{pmatrix}, \qquad y = (C_2, C_3) \begin{pmatrix} \nu_{12} \\ \nu_2 \end{pmatrix}.$$

It thus follows from Lemma 5.2 that $\binom{x_{12}}{x_2} \in L_\infty(\mathbb{R}^+)$, which of course implies that $x_2 \in L_\infty(\mathbb{R}^+)$. □

Another important instrument in the proof that we will establish is the following result.

LEMMA 5.4. *Consider the system $\dot{x} = Ax + Bu$, $x(0) = x_0$. Assume that (A, B) is controllable and $\sigma(A) \subset \mathbb{C}^- \cup \mathbb{C}^0$. Then for all $\varepsilon > 0$ there exists a control $u \in L_2(\mathbb{R}^+)$ such that $\int_0^\infty \|u(t)\|^2 \, dt < \varepsilon$ and $x(x_0, u)(t) \to 0 (t \to \infty)$.*

Proof. For the given system consider the fixed-endpoint regular LQ problem

$$\inf \left\{ \int_0^\infty \|u(t)\|^2 \, dt | u \in L_2(\mathbb{R}^+) \text{ and } x(x_0, u)(t) \to 0, \, t \to \infty \right\}.$$

It is well known (see also Theorem 4.1) that the above infimum is equal to $x_0^T K^+ x_0$, where K^+ is the maximal solution to the ARE: $A^T K + KA = KBB^T K$. We claim that $K^+ = 0$. Assume $K^+ \neq 0$. Since $K = 0$ is a solution to the ARE, we must have $0 \leqq K^+$. So, $K^+ \geqq 0$ and $K^+ \neq 0$. Consequently, there exists an orthogonal matrix S such that

$$SK^+S^T = \begin{pmatrix} K_1 & 0 \\ 0 & 0 \end{pmatrix},$$

with $K_1 > 0$. Denote $\bar{K} := SK^+S^T$, $\bar{A} := SAS^T$, $\bar{B} := SB$. Then we have $\bar{A}^T\bar{K} + \bar{K}\bar{A} = \bar{K}\bar{B}\bar{B}^T\bar{K}$. Decompose

$$\bar{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad \text{and} \quad \bar{B} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}.$$

It is easily seen that $A_{11}^T K_1 + K_1 A_{11} = K_1 B_1 B_1^T K_1$. Also, $K_1 A_{12} = 0$. Since $K_1 > 0$, this implies $A_{12} = 0$. Define $P := K_1^{-1}$. Then $P > 0$ and satisfies the Lyapunov equation $PA_{11}^T + A_{11}P = B_1 B_1^T$. Since (A_{11}, B_1) is controllable, this implies $\sigma(A_{11}) \subset \mathbb{C}^+$ (see, e.g., [21, Lemma 12.2]. This, however, contradicts the fact that $\sigma(A_{11}) \subset \sigma(\bar{A}) = \sigma(A) \subset \mathbb{C}^- \cup \mathbb{C}^0$. We conclude that the above infimum is zero. □

We have now collected the most important ingredients we need in the proof of our main theorem. In order to give this proof, we shall make a suitable direct sum

decomposition of the state space. Let K_f^+ be the solution of the ARE (3.1) defined by (5.3). Denote $A_f^+ := A - BR^{-1}(B^TK_f^+ + S)$. By Theorem 3.1 we have

$$X^+(A_f^+) = N,$$

$$X^0(A_f^+) = \ker \Delta,$$

$$X^-(A_f^+) = X^-(A^+) \cap \Delta^{-1}N^\perp.$$

Define $X_1 := X^+(A_f^+)$, $X_2 := X^0(A_f^+)$, and $X_3 := X^-(A_f^+)$. Then $\mathbb{R}^n = X_1 \oplus X_2 \oplus X_3$. Since X_1 is A^--invariant and since X_2 is also A^--invariant ($\ker \Delta = X^0(A_K)$ for *all* $K \in \Gamma$) we have

$$A^- = \begin{pmatrix} A_{11} & 0 & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix}, \tag{5.4}$$

for given matrices A_{ij}. We also have $K_f^+x = K^-x$ for all $x \in N$, and hence $A_f^+|X_1 = A^-|X_1$. Also, since $\ker \Delta \subset \Delta^{-1}N^\perp$ and therefore $\ker \Delta \subset \ker P_N$, for all $x \in \ker \Delta$ we have $K_f^+x = K^+x = K^-x$. Hence $A_f^+|X_2 = A^-|X_2$. Consequently,

$$A_f^+ = \begin{pmatrix} A_{11} & 0 & 0 \\ 0 & A_{22} & 0 \\ 0 & 0 & A'_{33} \end{pmatrix}, \tag{5.5}$$

for a given matrix A'_{33}. Note that $\sigma(A_{11}) \subset \mathbb{C}^+$, $\sigma(A_{22}) \subset \mathbb{C}^0$ and $\sigma(A'_{33}) \subset \mathbb{C}^-$. Since $X_1 \subset \ker K^-$ and K^- is symmetric,

$$K^- = \begin{pmatrix} 0 & 0 & 0 \\ 0 & K_{22}^- & K_{23}^- \\ 0 & K_{23}^{-T} & K_{33}^- \end{pmatrix}. \tag{5.6}$$

Furthermore, we claim that Δ has the form

$$\Delta = \begin{pmatrix} \Delta_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta_{33} \end{pmatrix}.$$

Indeed, by Theorem 3.1 we have $X_2 \oplus X_3 = \Delta^{-1}X_1^\perp$ and therefore we must have $\Delta_{13} = 0$. The other zero blocks are caused by the fact that $X_2 = \ker \Delta$ and by the symmetry of Δ. Combining the representations for K^- and Δ, we find

$$K^+ = \begin{pmatrix} K_{11}^+ & 0 & 0 \\ 0 & K_{22}^+ & K_{23}^+ \\ 0 & K_{23}^{+T} & K_{33}^+ \end{pmatrix}$$

for given matrices K_{ij}^+ (note that, in fact, $K_{23}^+ = K_{23}^-$ and $K_{22}^+ = K_{22}^-$). Combining all this, we find that

$$K_f^+ = \begin{pmatrix} 0 & 0 & 0 \\ 0 & K_{22}^- & K_{23}^- \\ 0 & K_{23}^{-T} & K_{33}^+ \end{pmatrix}. \tag{5.7}$$

We now proceed with the following lemma, which states that K_f^+ gives a lower bound for the optimal cost of the free-endpoint regular LQ problem.

LEMMA 5.5. *Assume that* (A, B) *is controllable,* $R>0$, *and* $\Gamma_{-}\neq\varnothing$. *For all* $x_0\in\mathbb{R}^n$ *and for all* $u\in U(x_0)$ *we have*

$$J(x_0,u)\geqq x_0^TK_f^+x_0+\int_0^\infty\|u(t)+R^{-1}(B^TK_f^++S)x(t)\|_R^2\,dt. \tag{5.8}$$

Here we have denoted $x(t):=x(x_0,u)(t)$.

Proof. Since $\Gamma_{-}\neq\varnothing$ we have $K^{-}\leqq 0$. Let $u\in U(x_0)$. It follows from Lemma 4.4 that $J(x_0,u)$ is either finite or equal to $+\infty$. Indeed, $J(x_0,u)=-\infty$ would imply $V_f^+(x_0)=-\infty$, which would contradict $\Gamma_{-}\neq\varnothing$. Of course, if $J(x_0,u)=+\infty$ then (5.8) holds trivially. Now assume that $J(x_0,u)$ is *finite*. By the fact that $K^{-}\leqq 0$ it follows from Lemma 4.3 that for all $T\geqq 0$

$$\int_0^T\|u(t)+R^{-1}(B^TK^{-}+S)x(t)\|_R^2\,dt\leqq J_T(x_0,u)-x_0^TK^{-}x_0.$$

Denote $v(t):=u(t)+R^{-1}(B^TK^{-}+S)x(t)$. It then follows that $\int_0^\infty\|v(t)\|_R^2\,dt<+\infty$, and hence that $v\in L_2(\mathbb{R}^+)$. Again using Lemma 4.3 and the fact that $-K^{-}\geqq 0$, we find that this implies $\lim_{T\to\infty}x^T(T)K^{-}x(T)$ exists (and is finite). Thus $K^{-}x$ must be bounded on $\mathbb{R}^+$. Denote $y(t):=K^{-}x(t)$. Since $\dot{x}=Ax+Bu$, we have that x, v, and y are related by the equations

$$\dot{x}=A^{-}x+Bv,\qquad y=K^{-}x.$$

Now let $\mathbb{R}^n$ be composed into $\mathbb{R}^n=X_1\oplus X_2\oplus X_3$ as introduced above. Write $K^{-}=(0,K_2^{-},K_3^{-})$, $B=(B_1^T,B_2^T,B_3^T)^T$, and $x=(x_1^T,x_2^T,x_3^T)^T$. Since $X_1=N$ is the undetectable subspace (with respect to $\mathbb{C}^{-}\cup\mathbb{C}^0$) of (K^{-},A^{-}), it is easily verified that the pair

$$\left((K_2^{-},K_3^{-}),\begin{pmatrix}A_{22}&A_{23}\\0&A_{33}\end{pmatrix}\right)$$

is detectable (with respect to $\mathbb{C}^{-}\cup\mathbb{C}^0$). Since $\sigma(A^{-})\subset\mathbb{C}^{+}\cup\mathbb{C}^0$ and since $X_2=X^0(A^{-})$, it can be verified that

$$\sigma\left(\begin{pmatrix}A_{11}&A_{13}\\0&A_{33}\end{pmatrix}\right)\subset\mathbb{C}^{+}.$$

Hence, $\sigma(A_{22})\subset\mathbb{C}^0$ and $\sigma(A_{33})\subset\mathbb{C}^{+}$. Also, we have

$$\begin{pmatrix}\dot{x}_2\\\dot{x}_3\end{pmatrix}=\begin{pmatrix}A_{22}&A_{23}\\0&A_{33}\end{pmatrix}\begin{pmatrix}x_2\\x_3\end{pmatrix}+\begin{pmatrix}B_2\\B_3\end{pmatrix}v,\qquad y=(K_2^{-},K_3^{-})\begin{pmatrix}x_2\\x_3\end{pmatrix}.$$

Since $v\in L_2(\mathbb{R}^+)$ and $y\in L_\infty(\mathbb{R}^+)$, by Lemma 5.3 (applied with $\mathbb{C}_g=\mathbb{C}^{-}\cup\mathbb{C}^0$) we have that $x_3\in L_\infty(\mathbb{R}^+)$.

Again by applying Lemma 4.3, this time with $K=K_f^+$, we find that for all $T\geqq 0$

$$J_T(x_0,u)=\int_0^T\|u(t)+R^{-1}(B^TK_f^++S)x(t)\|_R^2\,dt+x_0^TK_f^+x_0-x^T(T)K_f^+x(T). \tag{5.9}$$

Denote $w(t):=u(t)+R^{-1}(B^TK_f^++S)x(t)$. Combining (5.6), (5.7), and (5.9), we obtain that for all $T\geqq 0$

$$J_T(x_0,u)=\int_0^T\|w(t)\|_R^2\,dt+x_0^TK_f^+x_0-x_3^T(T)\Delta_{33}x_3(T)-x^T(T)K^{-}x(T). \tag{5.10}$$

Recall that $\lim_{T\to\infty}J_T(x_0,u)$ was assumed to be finite. Thus, $J_T(x_0,u)$ is a bounded function of T. Since also $x_3(T)$ and $x^T(T)K^{-}x(T)$ are bounded functions of T, (5.10) implies that $\int_0^\infty\|w(t)\|_R^2\,dt<\infty$. It follows that $w\in L_2(\mathbb{R}^+)$.

We again consider (5.10). Since now $J_T(x_0, u)$, $\int_0^T \|w(t)\|_R^2\, dt$ and $x^T(T)K^-x(T)$ *converge* for $T\to\infty$, it follows that $\lim_{T\to\infty} x_3^T(T)\Delta_{33}x_3(T)$ exists. Since $\Delta_{33}>0$ this implies that $\|x_3(T)\|$ converges as $T\to\infty$. Now, since $\dot{x}=Ax+Bu$, the variables x and w are related via $\dot{x}=A_f^+x+Bw$, and hence (see 5.5) $\dot{x}_3=A_{33}'x_3+B_3w$. Since $w\in L_2(\mathbb{R}^+)$ and $\sigma(A_{33}')\subset\mathbb{C}^-$ we have that $x_3\in L_2(\mathbb{R}^+)$. A fortiori, since $\|x_3(t)\|$ converges as $t\to\infty$, this yields $\lim_{t\to\infty} x_3(t)=0$. Using this, and the fact that $-K^-\geqq 0$, it then follows from (5.10) that (5.8) holds. □

Our next lemma states that, by choosing the control properly, the difference between K_f^+ and the value of the cost functional can be made arbitrarily small.

LEMMA 5.6. *Assume that* (A, B) *is controllable,* $R>0$, *and* $\Gamma\neq\varnothing$. *Then for all* $x_0\in\mathbb{R}^n$ *and for all* $\varepsilon>0$ *there exists an input* $u\in U(x_0)$ *such that* $J(x_0, u)\leqq x_0^TK_f^+x_0+\varepsilon$.

Proof. Again, let $\mathbb{R}^n$ be decomposed as above. It follows from (5.7) and (5.9) that for all $u\in L_{2,\mathrm{loc}}(\mathbb{R}^+)$ and for all $T\geqq 0$

$$(5.11)\quad J_T(x_0, u)=\int_0^T \|w(t)\|_R^2\, dt+x_0^TK_f^+x_0-(x_2^T(T), x_3^T(T))\begin{pmatrix} K_{22}^- & K_{23}^- \\ K_{23}^{-T} & K_{33}^+\end{pmatrix}\begin{pmatrix} x_2(T)\\ x_3(T)\end{pmatrix}.$$

Here, $w:=u+R^{-1}(B^TK_f^++S)x$. Since $\dot{x}=Ax+Bu$, the variables x and w are related by $\dot{x}=A_f^+x+Bw$, and hence (see (5.5))

$$\begin{pmatrix}\dot{x}_2\\ \dot{x}_3\end{pmatrix}=\begin{pmatrix} A_{22} & 0\\ 0 & A_{33}'\end{pmatrix}\begin{pmatrix} x_2\\ x_3\end{pmatrix}+\begin{pmatrix} B_2\\ B_3\end{pmatrix}w.$$

Note that $\sigma(A_{22})\subset\mathbb{C}^0$, $\sigma(A_{33}')\subset\mathbb{C}^-$ and that this system is controllable. Now let $\varepsilon>0$. It follows from Lemma 5.4 that there exists a control $w\in L_2(\mathbb{R}^+)$ such that $\int_0^\infty \|w(t)\|_R^2\, dt<\varepsilon$ and such that $x_2(T)\to 0$ and $x_3(T)\to 0$ as $T\to\infty$. Define $u:=-R^{-1}(B^TK_f^++S)x+w$. Then we have

$$J(x_0, u)=\lim_{T\to\infty} J_T(x_0, u)=\int_0^\infty \|w(t)\|_R^2\, dt+x_0^TK_f^+x_0\leqq\varepsilon+x_0^TK_f^+x_0.$$ □

We will now prove our main theorem.

Proof of Theorem 5.1. (i) This proof was already stated separately in Lemma 4.4.

(ii) Lemma 5.5 yields $J(x_0, u)\geqq x_0^TK_f^+x_0$ for all $u\in U(x_0)$. Together with Lemma 5.6 this implies $V_f^+(x_0)=x_0^TK_f^+x_0$ for all x_0.

(iii) Assume $\Gamma_-\neq\varnothing$. ($\Rightarrow$) Assume that for all x_0 there exists a control $u^*\in U(x_0)$ such that $J(x_0, u^*)=V_f^+(x_0)=x_0^TK_f^+x_0$. Let $x_0\in\mathbb{R}^n$ be arbitrary and let u^* be the corresponding optimal control. Denote $x^*:=x(x_0, u^*)$. By Lemma 5.5

$$x_0^TK_f^+x_0=J(x_0, u^*)\geqq x_0^TK_f^+x_0+\int_0^\infty \|u^*(t)+R^{-1}(B^TK_f^++S)x^*(t)\|_R^2\, dt.$$

It follows that u^* must be given by the feedback control law $u^*=-R^{-1}(B^TK_f^++S)x^*$. This implies that x^* satisfies the equation $\dot{x}^*=A_f^+x^*$. In terms of the decomposition introduced above, this of course yields $\dot{x}_2^*=A_{22}x_2^*$ and $\dot{x}_3^*=A_{33}'x_3^*$ (see 5.5). Since $\sigma(A_{33}')\subset\mathbb{C}^-$ we must have $x_3^*(t)\to 0 (t\to\infty)$. By (5.10)

$$J_T(x_0, u^*)=x_0^TK_f^+x_0-x_3^{*T}(T)\Delta_{33}x_3^*(T)-x^{*T}(T)K^-x^*(T).$$

By the fact that $J_T(x_0, u^*)\to x_0^TK_f^+x_0$ we obtain that $x^{*T}(T)K^-x^*(T)\to 0$ $(T\to\infty)$. Since K^- is semidefinite, a fortiori this implies $K^-x^*(T)\to 0$ $(T\to\infty)$. Using (5.6) this yields

$$K_{22}^-x_2^*(T)+K_{23}^-x_3^*(T)\to 0\qquad (T\to\infty).$$

Since $x_3^*(T) \to 0$ $(T \to \infty)$ the latter implies $K_{22}^- x_2^*(T) \to 0$ $(T \to \infty)$ or, equivalently, $K_{22}^- \exp(A_{22}T)x_2(0) \to 0$ $(T \to \infty)$. Now, $x_2(0)$ was completely arbitrary and therefore we find that

$$K_{22}^- e^{A_{22}T} \to 0 \qquad (T \to \infty).$$

Consequently, $K_{22}^-(Is - A_{22})^{-1}$ has all its poles in $\mathbb{C}^-$. On the other hand, however, since $\sigma(A_{22}) \subset \mathbb{C}^0$, it has all its poles in $\mathbb{C}^0$. Thus, $K_{22}^-(Is - A_{22})^{-1} = 0$, and hence $K_{22}^- = 0$. Since K^- is semidefinite this implies $K_{23}^- = 0$. It follows that $\ker \Delta = X_2 \subset \ker K^-$.

($\Leftarrow$) Conversely, assume $\ker \Delta \subset \ker K^-$. Then $K_{22}^- = 0$ and $K_{23}^- = 0$. Define $u = -R^{-1}(B^T K_f^+ + S)x$. We claim that this feedback law yields an optimal u. Indeed, by (5.11)

$$J_T(x_0, u) = x_0^T K_f^+ x_0 - x_3^T(T) K_{33}^+ x_3(T).$$

Moreover, $\dot{x}_3 = A_{33}' x_3$. Since $\sigma(A_{33}') \subset \mathbb{C}^-$ we have $x_3(T) \to 0$ $(T \to \infty)$. Thus $J(x_0, u) = x_0^T K_f^+ x_0 = V_f^+(x_0)$, so u is optimal.

(iv) The fact that $u^* = -R^{-1}(B^T K_f^+ + S)x^*$ is *unique* was already proven in (iii) ($\Rightarrow$). This concludes the proof of our theorem. □

Remark 5.7. At this point we would like to mention that, in addition to the option we have chosen in § 2, there is still another very natural and appealing way to formulate the regular LQ problem. Instead of restricting the class of controls to $U(x_0)$ in order to guarantee that the indefinite integrals in (2.2) are well-defined, it is also possible to choose $L_{2,\mathrm{loc}}(\mathbb{R}^+)$ for the class of admissible controls and to consider the following cost functional:

$$\tilde{J}(x_0, u) := \limsup_{T \to \infty} J_T(x_0, u).$$

Obviously, on the subclass $U(x_0) \subset L_{2,\mathrm{loc}}(\mathbb{R}^+)$ the functionals $\tilde{J}(x_0, \cdot)$ and $J(x_0, \cdot)$ coincide. Corresponding to this choice of cost functional, we can now consider the following version of the free-endpoint regular LQ problem:

$$\tilde{V}_f^+(x_0) := \inf\{\tilde{J}(x_0, u) | u \in L_{2,\mathrm{loc}}(\mathbb{R}^+)\}.$$

As it turns out, we can develop around this version of the problem a theory completely parallel to the one we developed in this section. In fact, Theorem 5.1 remains valid if in its statement we replace V_f^+ by $\tilde{V}_f^+$! In particular, both problems yield the same optimal controls u^*. Consequently, if u^* is optimal for the problem with functional $\tilde{J}(x_0, \cdot)$, then in fact $u^* \in U(x_0)$ and $\tilde{V}_f^+(x_0) = \tilde{J}(x_0, u^*) = \lim_{T \to \infty} J_T(x_0, u^*)$. Similar remarks hold for the fixed-endpoint problem.

6. Comparison and special cases. In this section we will discuss some questions that arise if we compare the optimal costs and optimal closed loop systems resulting from the free-endpoint and fixed-endpoint problem, respectively. In particular, we will establish conditions under which the respective optimal costs are the same. Also, conditions will be found under which the free-endpoint optimal closed loop system is asymptotically stable. Finally, we will show how our general results can be specialized to reobtain the most important results on the free-endpoint regular LQ problem with *positive semidefinite* cost functional. First, we have the following theorem.

THEOREM 6.1. *Assume that* (A, B) *is controllable,* $R > 0$, *and* $\Gamma \neq \emptyset$. *Then we have the following*:

(i) $K_f^+ = K^+$ *if and only if the pair* $(K^-, A - BR^{-1}S)$ *is detectable with respect to the stability set* $\mathbb{C}^- \cup \mathbb{C}^0$.

(ii) $\sigma(A_f^+) \subset \mathbb{C}^-$ *if and only if the pair* $(K^-, A - BR^-S)$ *is detectable with respect to* $\mathbb{C}^-$ *and* $\Delta > 0$.

Proof. (i) By (5.2), N is equal to the undetectable subspace of $(K^-, A - BR^{-1}S)$ with respect to $\mathbb{C}^- \cup \mathbb{C}^0$. Since K^+ is supported by the zero subspace, by Theorem 3.1 we have $K_f^+ = K^+$ if and only if $N = 0$.

(ii) ($\Leftarrow$) Detectability with respect to $\mathbb{C}^-$ implies detectability with respect to $\mathbb{C}^- \cup \mathbb{C}^0$. Hence $K_f^+ = K^+$ and $A_f^+ = A^+$. By [17, Thm. 5] $\Delta > 0$ if and only if $\sigma(A^+) \subset \mathbb{C}^-$.

($\Rightarrow$) Conversely, assume $\sigma(A_f^+) \subset \mathbb{C}^-$. By [17, Thm. 5] there is exactly one $K \in \Gamma$, namely $K = K^+$, such that $\sigma(A_K) \subset \mathbb{C}^- \cup \mathbb{C}^0$. Hence $K_f^+ = K^+$, $A_f^+ = A^+$. Consequently, $\Delta > 0$. Also, from (i) we obtain that the pair $(K^-, A - BR^{-1}S)$ is detectable with respect to $\mathbb{C}^- \cup \mathbb{C}^0$. Since $\Delta > 0$, $\sigma(A^-) \subset \mathbb{C}^+$. Hence $X^0(A^-) = 0$ so $(K^-, A - BR^{-1}S)$ is in fact detectable with respect to $\mathbb{C}^-$. □

We will now discuss how our results can be specialized to rederive some important "classical" results on the special case that the quadratic form ω is positive semidefinite. We have the following characterization of the positive semidefinite solutions of the ARE.

THEOREM 6.2. *Assume that* (A, B) *is controllable,* $R > 0$, $\Gamma_- \neq \emptyset$, *and* $\Gamma_+ \neq \emptyset$. *Let* $K \in \Gamma$ *be supported by* V. *Then* $K \in \Gamma_+$ *if and only if* $V \subset \ker K^-$.

Proof. By Theorem 3.1 we have $V \oplus \Delta^{-1}V^\perp = \mathbb{R}^n$.

($\Leftarrow$) Assume that $V \subset \ker K^-$. Then $\Delta^{-1}V^\perp = \{x \in \mathbb{R}^n \mid y^T K^+ x = 0, \text{ for all } y \in V\}$ and $K = K^+(I - P_V)$. Let $x \in \mathbb{R}^n$, $x = x_1 + x_2$ with $x_1 \in V$ and $x_2 \in \Delta^{-1}V^\perp$. It is easily seen that $x^T K x = x_2^T K^+ x_2$. Since $\Gamma_+ \neq \emptyset$ we have $K^+ \geqq 0$. It follows that $K \geqq 0$.

($\Leftarrow$) Conversely, if $K \geqq 0$ then for all $x \in V$ we have

$$0 \leqq x^T K x = x^T(K^- P_V + K^+(I - P_V))x = x^T K^- x.$$

Since $\Gamma_- \neq \emptyset$ we have $K^- \leqq 0$. It follows that $x^T K^- x = 0$, and hence that $x \in \ker K^-$. □

Our next result states that, under the assumption that $\Gamma_- \neq \emptyset$, if the ARE has positive semidefinite solutions at all, then it has a smallest positive semidefinite solution and this solution is equal to the one supported by N.

THEOREM 6.3. *Assume that* (A, B) *is controllable,* $R > 0$, *and* $\Gamma_- \neq \emptyset$. *Then the following hold: if* $\Gamma_+ \neq \emptyset$ *then* (i) $K_f^+ \in \Gamma_+$ *and* (ii) $K \in \Gamma_+$ *implies* $K_f^+ \leqq K$.

Proof. Since $N \subset \ker K^-$ it follows from Theorem 6.2 that $K_f^+ \in \Gamma_+$. Now assume $K \in \Gamma_+$ and K is supported by the A^--invariant subspace $V \subset X^+(A^-)$. Since $K \in \Gamma_+$ we have $V \subset \ker K^-$. Hence $V \subset \langle \ker K^- | A^- \rangle$ (the latter is the largest A^--invariant subspace in $\ker K^-$; see [21]). It follows that $V \subset N$. But then, by Theorem 3.2, $K_f^+ \leqq K$. □

From the above we deduce the following remarkable fact. Consider the free-endpoint regular LQ problem with *indefinite* cost functional. Let (A, B) be controllable. We already saw that the optimal cost is finite if we have $\Gamma_- \neq \emptyset$. Assume this to be the case. Then Theorem 6.3 states that *if the* ARE *has at least one positive semidefinite solution, then the optimal cost is given by the smallest of these solutions*! The case that the cost functional is positive semidefinite, i.e., $\omega(x, u) \geqq 0$, for all (x, u), is in fact a special case of this general principle. Indeed, if (A, B) is controllable and if $\omega \geqq 0$ then $\Gamma_+ \neq \emptyset$ (see [5]). Moreover, applying the latter to the controllable system $(-A, -B)$ and the same form $\omega \geqq 0$, we can also see that $\Gamma_- \neq \emptyset$. Thus we have reobtained Theorem 4.2(i).

Our next result shows that the fact that for the case $\omega \geqq 0$ optimal controls exist for all initial conditions is also a special case of a more general principle.

PROPOSITION 6.4. *Assume* (A, B) *is controllable,* $R > 0$, $\Gamma_- \neq \emptyset$, *and* $\Gamma_+ \neq \emptyset$. *Then* $\ker \Delta \subset \ker K^-$.

Proof. $\Gamma_- \neq \emptyset$ is equivalent to $K^- \leqq 0$ and $\Gamma_+ \neq \emptyset$ is equivalent to $K^+ \geqq 0$. Assume $x \in \ker \Delta$. Then $0 \leqq x^T K^+ x = x^T K^- x \leqq 0$. Thus $x^T K^- x = 0$, and hence $K^- x = 0$. □

By combining this with the above remarks and by applying Theorem 5.1(iii) and (iv) we reobtain Theorem 4.2(ii).

To conclude this section, we will briefly discuss what statements can be obtained from Theorem 6.1 for the case that our cost functional is positive semidefinite. In the rest of this section, assume that $\omega(x, u) \geqq 0$ for all (x, u). We claim that in this case

$$N = \langle \ker (Q - S^T R^{-1} S) | A - BR^{-1} S \rangle \cap X^+(A - BR^{-1} S). \tag{6.1}$$

First we claim that $\ker K^-$ is $(A - BR^{-1}S)$-invariant. Indeed, if $\omega \geqq 0$ then $Q - S^T R^{-1} S \geqq 0$. Also it is straightforward to verify that

$$(A - BR^{-1}S)^T K^- + K^-(A - BR^{-1}S) + Q - S^T R^{-1} S - K^- BR^{-1} B^T K = 0. \tag{6.2}$$

Let $x_0 \in \ker K^-$. Then from (6.2), $x_0^T (Q - S^T R^{-1} S) x_0 = 0$, and hence $(Q - S^T R^{-1} S) x_0 = 0$. Thus, again from (6.2), $K^-(A - BR^{-1}S)x_0 = 0$ so $(A - BR^{-1}S)x_0 \in \ker K^-$. It follows that $\langle \ker K^- | A - BR^{-1}S \rangle = \ker K^-$. Now, by using the interpretation of K^- as the optimal cost for a fixed-endpoint LQ problem in "reversed time" (see [21, Thm. 7]) it can be proved that

$$\begin{aligned} \ker K^- = \langle \ker (Q - S^T R^{-1} S) | A - BR^{-1} S \rangle \\ \cap (X^+(A - BR^{-1}S) \oplus X^0(A - BR^{-1}S)). \end{aligned} \tag{6.3}$$

Thus (6.1) follows immediately from (5.2). We have now shown that if $\omega \geqq 0$, then K_f^+ is in fact supported by the undetectable subspace of the pair $(Q - S^T R^{-1} S, A - BR^{-1} S)$ with respect to $\mathbb{C}^- \cup \mathbb{C}^0$. (See also [3, Thm. 1].) By applying Theorem 6.1(i) we may then conclude that $K_f^+ = K^+$ if and only if $(Q - S^T R^{-1} S, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^- \cup \mathbb{C}^0$ (see also [12, Cor., p. 356]).

Finally, we will re-establish the well-known fact that $\sigma(A_f^+) \subset \mathbb{C}^-$ if and only if $(Q - S^T R^{-1} S, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^-$ (see [6], [20], and [12]). Assume that $\omega \geqq 0$. We claim that if $(K^-, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^-$ then $\Delta > 0$. Indeed, if $(K^-, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^-$ then (K^-, A^-) is detectable with respect to $\mathbb{C}^-$. The latter is equivalent to

$$\langle \ker K^- | A^- \rangle \cap (X^+(A^-) \oplus X^0(A^-)) = 0. \tag{6.4}$$

By Theorem 3.1, $X^0(A^-) = \ker \Delta$. Also, since $\omega \geqq 0$, $\ker \Delta \subset \ker K^-$. Hence, by (6.4), $\ker \Delta + (\langle \ker K^- | A^- \rangle \cap X^+(A^-)) = 0$, whence $\ker \Delta = 0$. It follows that $\Delta > 0$. We may now conclude from Theorem 6.1(ii) that $\sigma(A_f^+) \subset \mathbb{C}^-$ if and only if the pair $(K^-, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^-$. From the fact that $\ker K^-$ is $(A - BR^{-1}S)$-invariant and from (6.3), the latter condition is, however, equivalent to the statement that the pair $(Q - S^T R^{-1} S, A - BR^{-1} S)$ is detectable with respect to $\mathbb{C}^-$.

7. Concluding remarks. In this paper we have studied just one of the many open basic questions that still exist in the context of linear quadratic optimal control. To name but a few of these open problems, we mention, for example, the question about the relationship between the *finite*-horizon free-endpoint problem and the infinite-horizon free-endpoint problem. It is well known that if the cost functional is positive semidefinite, then the finite-horizon optimal cost converges to the infinite-horizon optimal cost [1], [2], [9]. It would be interesting to investigate whether this is also true for the indefinite case. Another open problem is the *singular* LQ problem with indefinite cost functional, that is, the problem studied here without the assumption that R is positive definite. Recently [19] this problem was treated for the case that the

cost-functional is positive semidefinite. However, for both the free-endpoint case as well as the fixed-endpoint case, the indefinite version of this problem still remains to be solved.

Acknowledgments. I thank Dr. Jacob van der Woude and Professor Malo Hautus for some very useful discussions while the research leading to this paper was carried out.

REFERENCES

[1] B. D. O. ANDERSON AND J. B. MOORE, *Linear Optimal Control*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] F. M. CALLIER AND J. L. WILLEMS, *Criterion for the convergence of the solution of the Riccati differential equation*, IEEE Trans. Automat. Control, 26 (1981), pp. 1232-1242.

[4] W. A. COPPEL, *Matrix quadratic equation*, Bull. Austral. Math. Soc., 10 (1974), pp. 377-401.

[5] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102-199.

[6] ———, *When is a linear control system optimal*? Trans. ASME J. Basic Engrg., 83 (1964), pp. 51-60.

[7] V. KUCERA, *A contribution to matrix quadratic equations*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 344-347.

[8] ———, *On non-negative definite solutions to matrix quadratic equations*, Automatica, 8 (1972), pp. 413-423.

[9] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[10] K. MÅRTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17-49.

[11] B. P. MOLINARI, *Conditions for nonpositive solutions of the linear matrix inequality*, IEEE Trans. Automat. Control, 20 (1975), pp. 804-806.

[12] ———, *The time-invariant linear-quadratic optimal control problem*, Automatica, 13 (1977), pp. 347-357.

[13] P. J. MOYLAN, *On a frequency domain condition in linear optimal control theory*, IEEE Trans. Automat. Control, 29 (1975), p. 806.

[14] V. M. POPOV, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roumaine Sci. Tech. Ser. Electrotech. Energet., 9 (1964), pp. 629-690.

[15] J. M. SCHUMACHER, *Dynamic Feedback in Finite and Infinite Dimensional Linear Systems*, Math. Centre Tracts, 143, Amsterdam, the Netherlands, 1981.

[16] M. SHAYMAN, *Geometry of the algebraic Riccati equation—part* 1, SIAM J. Control Optim., 21 (1983), pp. 375-393.

[17] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621-634.

[18] ———, *On the existence of a non-positive solution to the Riccati equation*, IEEE Trans. Automat. Control, 19 (1974), pp. 592-593.

[19] J. C. WILLEMS, A. KITAPÇI, AND L. M. SILVERMAN, *Singular optimal control, a geometric approach*, SIAM J. Control Optim., 24 (1986), pp. 323-337.

[20] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control Optim., 6 (1968), pp. 681-698.

[21] ———, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 43-52, January 1989

004

SUBSPACE INVARIANCE OF BROYDEN'S θ-CLASS AND ITS APPLICATION TO NONLINEAR CONSTRAINED OPTIMIZATION*

RODRIGO FONTECILLA†

Abstract. Broyden's θ-class of updating formulae has the following property. If the current approximation matrix is symmetric and positive definite, then the updated matrix maintains those same properties under certain conditions. It is shown that if the current approximation matrix is symmetric and positive definite on a subspace of $\mathbf{R}^n$, then the updated matrix is symmetric and positive definite along the same subspace. An application of this result to the implementation of a quasi-Newton method for solving nonlinear constrained optimization problems is presented.

Key words. unconstrained minimization, constrained minimization, quasi-Newton methods, Broyden's θ-class

AMS(MOS) subject classifications. 65K05, 65K10

1. Introduction.

Consider the unconstrained minimization problem

$$\underset{x\in\mathbf{R}^n}{\text{minimize}}\, f(x), \tag{1.1}$$

with solution x_* for a twice continuously differentiable function $f:\mathbf{R}^n\to\mathbf{R}$ with gradient $\nabla f(x)$ and Hessian $\nabla^2 f(x)$. This problem is often solved by a class of quasi-Newton methods that generate points $\{x\}$ and $n\times n$ matrices $\{B\}$ such that

$$x_+ = x-\alpha d, \qquad d = B^{-1}\nabla f(x). \tag{1.2}$$

The stepsize $\alpha>0$ is chosen such that $f(x_+)<f(x)$. In order to ensure that $-d$ is a descent direction for f at x, we require that all B be symmetric and positive definite. In addition, we ask that the secant equation

$$B_+ s = y \tag{1.3}$$

be satisfied, where

$$s = x_+ - x \quad\text{and}\quad y = \nabla f(x_+)-\nabla f(x). \tag{1.4}$$

A general class of methods for computing a symmetric, positive definite B_+ satisfying (1.3) from a symmetric and positive definite B is Broyden's bounded θ-class of methods, which is given by an update formula depending on a constant θ which may even vary throughout the process. The matrices B approximating the Hessian $\nabla^2 f(x_*)$ are given by

$$B_+ = U(s,y,B,\theta) = B-\frac{Bss^TB}{s^TBs}+\frac{yy^T}{y^Ts}+\theta ww^T, \tag{1.5a}$$

with

$$w=\sqrt{s^TBs}\left[\frac{y}{y^Ts}-\frac{Bs}{s^TBs}\right]. \tag{1.5b}$$

Any $U(s,y,B,\theta)$ with $\theta\in[0,1]$ belongs to the "convex hull" of the differentiable function problem (DFP) and BFGS formulae, which are obtained as special cases, namely, $\theta=1$ for the DFP, and $\theta=0$ for the BFGS. Standard conditions are the

* Received by the editors July 21, 1986; accepted for publication (in revised form) April 4, 1988.

† Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, Maryland 20742.

following:

(A1) The functional $f \in C^2(D)$, with D an open convex subset of $\mathbf{R}^n$.

(A2) The point x_* is in D, and is a local minimizer of f with $\nabla f(x_*) = 0$ and $\nabla^2 f(x_*)$ positive definite.

(A3) The Hessian $\nabla^2 f(x)$ is Lipschitz continuous on D.

It has been shown [2], [3], [10] that under conditions (A1)-(A3), for any $x_0 \in D$ sufficiently close to x_*, and any positive definite matrix B_0 close to $\nabla^2 f(x_*)$, these methods generate a sequence of points $\{x_k\}$ and positive definite matrices $\{B_k\}$ with

$$\sup \|B_k\| < \infty, \qquad \sup \|B_k^{-1}\| < \infty, \tag{1.6}$$

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = 0 \quad \text{if } x_k \neq x_*. \tag{1.7}$$

This means that the sequence $\{x_k\}$ converges q-superlinearly to x_*. This convergence result rests heavily on the fact that if B is positive definite and $s^T y > 0$, then B_+ is also positive definite. However, this property does not necessarily hold if B is positive definite only on a subspace of $\mathbf{R}^n$. Such a situation arises in solving nonlinear constrained optimization problems; that is, the Hessian we approximate is symmetric and positive definite only on a subspace of $\mathbf{R}^n$. Since the Hessian is known to have such properties, we want to maintain the same properties for the approximating matrices B. We shall give conditions to guarantee that the matrices B are positive definite on a subspace of $\mathbf{R}^n$.

We present our work in the following fashion. In § 2, we present our main result. We prove that methods in Broyden's bounded θ-class satisfy the following property. If the current approximation matrix B is positive definite in the current subspace $S \in \mathbf{R}^n$, if the step s is in S, and if $y^T s > 0$, then the updated matrix B_+ is positive definite in S. In § 3, we show that the matrices B still satisfy a bounded deterioration relation. In § 4, we apply this result to solving nonlinear constrained optimization problems with some numerical experiments on a set of 10 test problems.

2. Hereditary positive definiteness. In this section we show that if the matrix B is positive definite on a subspace S of $\mathbf{R}^n$, and if $s \in S$ and $y^T s > 0$, then B_+ generated by (1.5) is positive definite in S, even when y, s are not necessarily given by (1.4).

In the following, we want the subspace $S(x)$ and the $n \times m$ full rank matrix $N(x)$ to be related by

$$S(x) = \{y \in \mathbf{R}^n \colon N(x)^T y = 0\}, \tag{2.1}$$

$$S(x)^{\perp} = \{N(x)z \colon z \in \mathbf{R}^m\}. \tag{2.2}$$

Thus, the orthogonal projection onto $S(x)$ is given by

$$P(x) = I - N(x)[N(x)^T N(x)]^{-1} N(x)^T. \tag{2.3}$$

In the following, to ease the notation, arguments for S, N, and P will be deleted, so $S(x)$, $N(x)$, and $P(x)$ become S, N, and P, respectively.

LEMMA 2.1. *Let A be an $n \times n$ symmetric matrix. Then A is positive definite on S if and only if $PAP + NN^T$ is positive definite on all of $\mathbf{R}^n$. Moreover, if A is positive definite on S, then the matrix $PA + NN^T$ is nonsingular.*

Proof. (i) Let A be a positive definite matrix on S, and let $x \in \mathbf{R}^n$, $x \neq 0$. Then $x = y + w$ for some $y \in S$ and $w = N$, $z \in S^{\perp}$. At least one of w and y must be nonzero. Since $Py = y$, $Pw = 0$, and $N^T y = 0$, we get

$$x^T(PAP + NN^T)x = y^T A y + \|N^T w\|_2^2 = y^T A y + \|N^T N z\|_2^2.$$

Thus, the right-hand side is strictly positive unless $y = 0$ and $z = 0$, but if $y = 0$ then $w \neq 0$, so $z \neq 0$. It follows that $PAP + NN^T$ is positive definite.

(ii) Now let $PAP + NN^T$ be positive definite and $x \in S$, $x \neq 0$; then

$$x^T A x = x^T(PAP + NN^T)x > 0,$$

so A is positive definite on S.

To prove the second part, suppose that $x \in \mathbf{R}^n$, $x \neq 0$, and

$$(PA + NN^T)x = 0.$$

Then $PAx = -NN^T x$, and since $PAx \in S$, this implies that $NN^T x = 0$ and $PAx = 0$. Hence, $N^T x = 0$, so $x \in S$ and $Px = x$. But $PAx = 0$, so $0 = x^T(PAx) = x^T A x$, which implies that A is not positive definite on S. □

LEMMA 2.2. *Let A be an $n \times n$ symmetric matrix and positive definite along S. Then, there exists a positive constant $\bar{c}$ such that the matrix $A + cNN^T$ is positive definite for all $c > \bar{c}$.*

Proof. The proof can be found in [1, Lemma 1.25]. □

The following result relies heavily on the hereditary positive definite property of the methods (1.5a) and (1.5b).

THEOREM 2.3. *Let the matrix $B \in \mathbf{R}^{n \times n}$ and vectors y, s be such that B is symmetric and positive definite on a subspace S of $\mathbf{R}^n$, $y^T s > 0$, and $s \in S$. Then, the matrix B_+ generated by the update formula* (1.5) *is well defined and is symmetric and positive definite on S.*

Proof. If we multiply both sides of (1.5) by P, and since $s \in S$, we get

$$PB_+P = U(s, Py, PBP, \theta), \tag{2.4}$$

$$PB_+P + NN^T = U(s, Py, PBP + NN^T, \theta). \tag{2.5}$$

Since $PBP + NN^T$ is positive definite and $s^T(Py) = s^T y > 0$, we have that $PB_+P + NN^T$ is positive definite, and from Lemma 2.1 we get that B_+ is positive definite on S. □

3. Preserving bounded deterioration. In this section we show that the matrices B generated by the update formula (1.5) preserve a bounded deterioration property even when they are not positive definite. This result is fundamental to proving any local convergence result related to the use of these matrices and the update formula (1.5) (see [2] and [3]).

It is necessary to exhibit a bounded deterioration property for the matrices B in order to show that these matrices, which are approximating a Hessian although they do not converge to it, tend not to go too far away from it. In other words, the approximation to the Hessian deteriorates, but in a bounded fashion.

For the remainder of this section let x_* be a point in $\mathbf{R}^n$ and let $H(x_*) = H_*$ be a symmetric matrix that is positive definite along $S(x_*) = S_*$. Let σ be defined by

$$\sigma = \max\{\|x + s - x_*\|, \|x - x_*\|\}.$$

Hereafter, we will assume that y and s are nonzero vectors. To be able to use Theorem 2.3 we need conditions that will ensure $y^T s > 0$. The following result shows when this is possible.

THEOREM 3.1. *Assume σ, the vectors y and s, and the matrices N and H_* satisfy*

$$\frac{\|y - H_* s\|}{\|s\|} = O(\sigma), \tag{3.1}$$

$$\|N_* N_*^T - NN^T\| = O(\sigma), \tag{3.2}$$

and $s \in S$. Then there exists a positive constant ε such that if $\sigma \leqq \varepsilon$, then $y^T s > 0$.

Proof. Since H_* is positive definite along S_* from Lemma 2.2 there exists a positive constant $\bar{c}$ such that the matrix $H_* + cN_* N_*^T$ is positive definite for all $c > \bar{c}$. Consider

$$M^{-2} = H_* + cN_* N_*^T \quad \text{for } c > \bar{c}.$$

If $s \in S$ then $N^T s = 0$, so

$$\begin{aligned} \|My - M^{-1}s\| &\leqq \|M\| \, \|y - M^{-2}s\| \\ &\leqq \|M\| [\|y - H_* s\| + c\|N_* N_*^T - NN^T\| \, \|s\|] \\ &\leqq O(\sigma)\|M^{-1}s\|. \end{aligned} \tag{3.3}$$

Consider

$$y^T s = (My)^T (M^{-1}s) = (My - M^{-1}s)^T (M^{-1}s) + (M^{-1}s)^T (M^{-1}s).$$

Taking norms and using the Cauchy-Schwartz inequality, we get

$$1 - \frac{\|My - M^{-1}s\|}{\|M^{-1}s\|} \leqq \frac{y^T s}{\|M^{-1}s\|^2} \leqq 1 + \frac{\|My - M^{-1}s\|}{\|M^{-1}s\|}.$$

Hence, by using (3.3) we can now choose ε sufficiently small and obtain our desired result. □

In general, it is not difficult to satisfy (3.1) or (3.2). It is usually true that $N(x)$ is twice continuously differentiable and therefore (3.2) holds.

The following is a technical lemma used later in the section.

LEMMA 3.2. *Let the matrices B' and B_+ be defined by*

$$B' = U(z, z, B, \theta) \quad \text{and} \quad B_+ = U(z, y, B, \theta),$$

and assume the vectors y and z satisfy $z^T B z \neq 0$ and

$$\frac{\|y - z\|}{\|z\|} = O(\sigma). \tag{3.4}$$

Then there exists $\varepsilon > 0$ such that if $\sigma \leqq \varepsilon$,

$$\|B_+ - B'\|_F \leqq [\|B\|_F + O(1)]O(\sigma). \tag{3.5}$$

Proof. Consider

$$y^T z = (y - z)^T z + z^T z.$$

Taking norms and using the Cauchy-Schwartz inequality, we obtain

$$1 - \frac{\|y - z\|}{\|z\|} \leqq \frac{y^T z}{\|z\|^2} \leqq 1 + \frac{\|y - z\|}{\|z\|}.$$

We now choose ε sufficiently small so that (3.4) implies

$$1 - O(\sigma) \leqq \frac{y^T z}{\|z\|^2} \leqq 1 + O(\sigma)$$

and $y^T z > 0$. We can restrict ε further so that $\frac{1}{2} \leqq 1 - O(\sigma)$ and therefore $y^T z \geqq \|z\|^2/2$. Now consider

$$B_+ - B' = E_1 + \theta(E_2 + E_2^T) + \theta z^T B z E_3,$$

where

$$E_1 = \frac{yy^T}{y^T z} - \frac{zz^T}{z^T z},$$

$$E_2 = \left[\frac{y}{y^T z} - \frac{z}{z^T z}\right](Bz)^T,$$

$$E_3 = \frac{yy^T}{(y^T z)^2} - \frac{zz^T}{(z^T z)^2}.$$

Rewrite E_1 as

$$E_1 = \frac{y(y-z)^T + (y-z)z^T}{y^T z} + zz^T\left[\frac{(z-y)^T z}{(y^T z)(z^T z)}\right].$$

Recall that $y^T z \geqq \|z\|^2/2$ and that (3.4) yields $\|y\| \leqq (1 + O(\sigma))\|z\|$. Taking norms, and using the triangle inequality and (3.4), we obtain

$$\|E_1\|_F \leqq 2\frac{\|y\|}{\|z\|}\frac{\|y-z\|}{\|z\|} + 4\frac{\|y-z\|}{\|z\|} = O(\sigma).$$

By a similar argument we obtain

$$\|E_2\|_F \leqq 4\frac{\|y-z\|}{\|z\|^2}\|Bz\| \leqq \|B\|O(\sigma),$$

$$|z^T B z|\,\|E_3\|_F \leqq \|z\|^2\|B\|\frac{\|E_1\|_F}{\|z\|^2} = \|B\|O(\sigma).$$

Therefore, we get

$$\|B_+ - B'\|_F \leqq O(\sigma) + 2\theta\|B\|O(\sigma) + \theta\|B\|O(\sigma)$$

which yields (3.5). □

The following result was given by Griewank and Toint [7] for the positive definite case. We now show that the matrices B still satisfy a bounded deterioration relation.

THEOREM 3.3. *Let the assumptions of Theorem* 3.1 *hold and the matrix B be symmetric and positive definite on S. Then there exist $\varepsilon > 0$ such that if $\sigma \leqq \varepsilon$, the matrix B_+ given by* (1.5) *satisfies*

$$\|B_+ - H_*\|_M \leqq [1 + O(\sigma)]\|B - H_*\|_M + O(\sigma), \tag{3.6}$$

with $\|Q\|_M = \|MQM\|_F$ and M as defined in the proof of Theorem 3.1.

Proof. First note that using Theorem 3.1 we can choose ε sufficiently small so that $y^T s > 0$. Thus, by Theorem 2.3 we obtain that B_+ is positive definite on S. Since H_* is positive definite on S_*, there exists a positive constant c_1 such that $H_* + c_1 N_* N_*^T$ is positive definite; since the matrices B_+ and B are positive definite on S, there exist positive constants c_2 and c_3 such that $B + c_2 NN^T$ and $B_+ + c_3 NN^T$ are positive definite. Now choose $c = \max\{c_1, c_2, c_3\}$, define M as in Theorem 3.1, i.e.,

$$M^{-2} = H_* + cN_* N_*^T = H_*^c,$$

and consider the positive definite matrices

$$B_c = B + cNN^T \quad \text{and} \quad B_+^c = B_+ + cNN^T. \tag{3.7}$$

Since $s \in S$, we have

$$\|y - M^{-2}s\| = \|y - H_* s - c(N_* N_*^T - NN^T)s\|.$$

The triangle inequality, (3.1), and (3.2) yield

$$\frac{\|y - z\|}{\|z\|} = O(\sigma), \tag{3.8}$$

with $z = M^{-2}s$. Let $\bar{B}$ be given by (1.5) with z instead of y:

$$\bar{B} = U(s, z, B, \theta),$$

and let $B_c' = \bar{B} + cNN^T = U(s, z, B_c, \theta)$.

For any nonsingular symmetric matrix M, (1.5) yields

$$MB_c'M = U(M^{-1}s, Mz, MB_cM, \theta) = U(M^{-1}s, M^{-1}s, MB_cM, \theta), \tag{3.9}$$

$$MB_+^cM = U(M^{-1}s, My, MB_cM, \theta). \tag{3.10}$$

Now use Lemma 3.2 with $M^{-1}s$, My, $MB_c'M$, and MB_cM in place of z, y, B', and B, respectively. Note that (3.4) holds because of (3.8) and $(M^{-1}s)^T(MB_cM)(M^{-1}s) \neq 0$ because $s \in S$. Thus, using Lemma 3.2 and further restricting ε if necessary, we obtain

$$\begin{aligned}\|MB_+^cM - MB_c'M\|_F &\leqq [\|MB_cM\|_F + O(1)]O(\sigma) \\ &\leqq [\|MB_cM - I\|_F + O(1)]O(\sigma).\end{aligned}$$

Therefore, we have

$$\|B_+^c - B_c'\|_M \leqq [\|B_c - H_*^c\|_M + O(1)]O(\sigma). \tag{3.11}$$

When we use the fact that $\|Q\|_F^2 = \operatorname{tr}(Q^TQ)$ and $z = M^{-2}s$ a tedious but elementary calculation using (3.9) yields

$$\begin{aligned}&\|B_c' - H_*^c\|_M^2 - \|B_c - H_*^c\|_M^2 \\ &\quad = \|MB_c'M - I\|_F^2 - \|MB_cM - I\|_F^2 \\ &\quad = -(1-\theta)\left\{\left(1 - \frac{s^TB_cM^2B_cs}{s^TB_cs}\right)^2 + 2\left[\frac{s^TB_cM^2B_cM^2B_cs}{s^TB_cs} - \left(\frac{s^TB_cM^2B_cs}{s^TB_cs}\right)^2\right]\right\} \\ &\qquad - \theta\left\{\left(1 - \frac{s^TB_cs}{s^Ts}\right)^2 + 2\theta\left[\frac{s^TB_cM^2B_cs}{s^Ts} - \left(\frac{s^TB_cs}{s^Ts}\right)^2\right]\right\} \\ &\qquad - \theta(1-\theta)\left\{\left(\frac{s^TB_cM^2B_cs}{s^TB_cs}\right)^2 - \left(\frac{s^TB_cs}{s^Ts}\right)^2\right\}.\end{aligned}$$

Now, using the Schwartz inequality, we can easily show that all three brackets are nonnegative and since $\theta \in [0, 1]$ we get

$$\|B_c' - H_*^c\|_M \leqq \|B_c - H_*^c\|_M. \tag{3.12}$$

Now (3.11), (3.12), and the triangle inequality give

$$\begin{aligned}\|B_+^c - H_*^c\|_M &\leqq \|B_+^c - B_c'\|_M + \|B_c' - H_*^c\|_M \\ &\leqq [\|B_c - H_*^c\|_M + O(1)]O(\sigma) + \|B_c - H_*^c\|_M \\ &\leqq [1 + O(\sigma)]\|B_c - H_*^c\|_M + O(\sigma).\end{aligned} \tag{3.13}$$

Finally, using (3.7), (3.2), and the triangle inequality, we obtain our desired result (3.6). □

4. An application to nonlinear constrained optimization. Consider the following nonlinear constrained optimization problem (NLCOP):

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g_i(x)=0, \quad i=1,\cdots,m, \end{aligned} \tag{4.1}$$

where f and g_i are functionals defined on an open convex subset D of $\mathbf{R}^n$, twice continuously differentiable with second derivatives that are Lipschitz continuous in a neighborhood of the solution x_*. The solution of (4.1) is based on the use of the Lagrangian function defined as

$$l(x,\lambda)=f(x)+g(x)^T\lambda, \tag{4.2}$$

with $g=(g_1,\cdots,g_m)^T$ and $\lambda\in\mathbf{R}^m$ being the Lagrange multipliers. The second-order sufficiency conditions state that if x_* is a regular point (i.e., $\nabla g(x_*)$ full rank), with $g(x_*)=0$ and Lagrange multipliers λ_* such that $\nabla_x l(x_*,\lambda_*)=0$ and for all $z\in\mathbf{R}^n$ with $\nabla g_*^T z=0$, $z^T\nabla_x^2 l(x_*,\lambda_*)z>0$, then x_* is a minimizer of (4.1).

Hence, our problem can be restated as follows. Find (x_*,λ_*) such that $\nabla l(x_*,\lambda_*)=0$ with the Hessian $\nabla_x^2 l(x_*,\lambda_*)$ being positive definite on a subspace of $\mathbf{R}^n$ given by

$$S(x_*)=S_*=\{z\in\mathbf{R}^n:\nabla g(x_*)^T z=0\}. \tag{4.3}$$

A method initially proposed by Tapia [11] and more recently studied by Fontecilla [5] is based on solution of the following nonlinear problem:

$$h(x)=P(x)\nabla f(x)+\nabla g(x)g(x)=0, \tag{4.4}$$

with $P(x)=I-\nabla g(x)[\nabla g(x)^T\nabla g(x)]^{-1}\nabla g(x)^T$. It is easily seen that $h(x_*)=0$ if and only if $\nabla l(x_*,\lambda_*)=0$ with $\lambda_*=-(\nabla g_*^T\nabla g_*)^{-1}\nabla g_*^T\nabla f_*$. The Jacobian of h at the solution x_* is $P_*\nabla_x^2 l(x_*,\lambda_*)+\nabla g_*\nabla g_*^T$, which by Lemma 2.1 is nonsingular. We define a Newton-like method for solving (4.4) as follows.

ALGORITHM 4.5.

Step 1. Given x_0

(4.5a) Step 2. Set $\lambda_{k+1}=-(\nabla g_k^T\nabla g_k)^{-1}\nabla g_k^T\nabla f_k$

(4.5b) Step 3. Solve $[P_k\nabla_x^2 l(x_k,\lambda_{k+1})+\nabla g_k\nabla g_k^T]s_k=-h_k$

(4.5c) Step 4. Set $x_{k+1}=x_k+s_k$

Step 5. Set $k=k+1$. Go to Step 2.

If second-order information is unavailable or too expensive to evaluate, the Hessian $H_*=\nabla_x^2 l(x_*,\lambda_*)$ can be approximated using matrices generated with the update formula (1.5). The algorithm follows.

ALGORITHM 4.6.

Step 1. Given x_0, B_0

(4.6a) Step 2. Set $\lambda_{k+1}=-(\nabla g_k^T\nabla g_k)^{-1}\nabla g_k^T\nabla f_k$

(4.6b) Step 3. Solve $(P_kB_k+\nabla g_k\nabla g_k^T)s_k=-(P_k\nabla f_k+\nabla g_k g_k)$

(4.6c) Step 4. Set $w_k=P_ks_k$

(4.6d) Step 5. Set $y_k=\nabla_x l(x_k+w_k,\lambda_{k+1})-\nabla_x l(x_k,\lambda_{k+1})$

(4.6e) Step 6. Set $B_{k+1}=U(w_k,y_k,B_k,\theta_k)$

(4.6f) Step 7. Set $x_{k+1}=x_k+s_k$

Step 8. Set $k=k+1$. Go to Step 2.

Note that if B_k is positive definite on S_k the linear system (4.6b) has a unique solution. Further, note that the step used to update the matrices B_k is $w_k \in S_k$ and that s_k is the step taken from x_k to get x_{k+1}. In [6] we showed that Algorithm 4.6 generates a sequence $\{x_k\}$ converging to x_* two-step q-superlinearly.

Algorithm 4.6 with the BFGS update in (4.6e) was run on a set of 10 test problems. The program was written using MATLAB, an interactive computer program for linear algebra computations developed by Cleve Moler. The double-precision version of MATLAB was run on a VAX 11/780 running UNIX™/4.2. The test problems were taken from Hock and Schittkowski [8] and Nocedal and Overton [9]. On all of them, the Hessian of the Lagrangian ($\nabla_x^2 l(x_*, \lambda_*)$) is not positive definite.

We were interested in studying the behavior of the eigenvalues of the matrices B_k and $P_k B_k P_k + \nabla g_k \nabla g_k^T$ at each iteration, and comparing them with the eigenvalues of the matrices $H_* = \nabla_x^2 l(x_*, \lambda_*)$ and $P_* H_* P_* + \nabla g_* \nabla g_*^T$, respectively. We ran the test problems with two different initial matrices B_0. We first tried a finite-difference (FD) approximation of $\nabla_x^2 l(x_0, \lambda_0)$ with a stepsize $h_j = \max(10^{-6}, 10^{-6}|x_0^{(j)}|)$ for the jth column and where $x_0^{(j)}$ is the jth component of x_0. We then checked that B_0 was positive definite along S_0 by verifying that the eigenvalues of $P_0 B_0 P_0 + \nabla g_0 \nabla g_0^T$ were all positive. The second choice for B_0 was the identity.

We stopped the algorithm when $\|h(x_k)\| = \|P_k \nabla f_k + \nabla g_k g_k\|$ and when $\|g_k\|$ became less than 10^{-6}. We also stopped if either the maximum number of iterations (50) was reached or the condition number of $P_k B_k + \nabla g_k \nabla g_k^T$ became greater than 10^8.

The results are summarized in Tables 1.1 and 1.2 for $B_0 = FD[\nabla_x^2 l(x_0, \lambda_0)]$ and in Tables 2.1 and 2.2 for $B_0 = I$. The first row of the tables gives the triplet (TP, n, m) where TP is the number of the test problem as given in Hock and Schittkowski [8]; if TP is P1 or P2 they correspond to those given by Nocedal and Overton [9]. The number of variables is n, and m is the number of constraints. The first column in Table 1.1 is read as follows:

eig (H_*)	the eigenvalues of H_*;
eig ($P_* H_*$)	the eigenvalues of $P_* H_* P_* + \nabla g_* \nabla g_*^T$;
eig (B_0)	the eigenvalues of B_0;
eig ($P_0 B_0$)	the eigenvalues of $P_0 B_0 P_0 + \nabla g_0 \nabla g_0^T$;
eig (B_k)	the eigenvalues of B_k at the last iteration;
eig ($P_k B_k$)	the eigenvalues of $P_k B_k P_k + \nabla g_k \nabla g_k^T$ at the last iteration.

The results in the tables are given as follows. We give a triplet (neg, zero, pos),

TABLE 1.1
B_0 using finite differences.

(TP, n, m)	(6, 2, 1)	(26, 3, 1)	(39, 4, 2)	(46, 5, 2)	(47, 5, 3)
eig (H_*)	(0, 1, 1)	(0, 2, 1)	(0, 1, 3)	(0, 1, 4)	(0, 1, 4)
eig ($P_* H_*$)	(0, 0, 2)	(0, 1, 2)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
eig (B_0)	(0, 1, 1)	(0, 2, 1)	(0, 1, 3)	(0, 1, 4)	(0, 0, 5)
eig ($P_0 B_0$)	(0, 0, 2)	(0, 1, 2)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
eig (B_k)	(0, 1, 1)	*	(0, 1, 3)	(0, 1, 4)	(0, 0, 5)
eig ($P_k B_k$)	(0, 0, 2)	*	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
id	0.9899	0.4379	0.3873	0.4472	0.2236
fh	1.5×10^{-7}	*	4.2×10^{-7}	3.8×10^{-8}	1.3×10^{-8}
fg	6.9×10^{-9}	*	3×10^{-10}	2.7×10^{-13}	1.5×10^{-12}
iter	7	*	12	6	5

* No convergence. $P_* H_* P_* + \nabla g_* \nabla g_*^T$ singular to machine precision.

TABLE 1.2
B_0 using finite differences.

(TP, n, m)	(56, 7, 4)	(78, 5, 3)	(104, 8, 4)	($P1$, 2, 1)	($P2$, 3, 2)
eig (H_*)	(2, 0, 5)	(1, 0, 4)	(2, 0, 6)	(1, 0, 1)	(2, 0, 1)
eig (P_*H_*)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
eig (B_0)	(2, 0, 5)	(1, 0, 4)	(2, 0, 6)	(1, 0, 1)	(2, 0, 1)
eig (P_0B_0)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
eig (B_k)	(2, 0, 5)	(1, 0, 4)	(2, 0, 6)	(1, 0, 1)	(2, 0, 1)
eig (P_kB_k)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
id	1.1736	0.3965	0.1439	0.24	0.7533
fh	4.4×10^{-7}	6.2×10^{-10}	6.3×10^{-7}	4.5×10^{-7}	7.1×10^{-11}
fg	8×10^{-8}	4.3×10^{-13}	2.4×10^{-9}	2.3×10^{-8}	7×10^{-12}
iter	10	6	6	5	5

TABLE 2.1
B_0 using the identity.

(TP, n, m)	(6, 2, 1)	(26, 3, 1)	(39, 4, 2)	(46, 5, 2)	(47, 5, 3)
eig (B_0)	(0, 0, 2)	(0, 0, 3)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
eig (P_0B_0)	(0, 0, 2)	(0, 0, 3)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
eig (B_k)	(0, 0, 2)	(0, 0, 3)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
eig (P_kB_k)	(0, 0, 2)	(0, 0, 2)	(0, 0, 4)	(0, 0, 5)	(0, 0, 5)
fh	1.5×10^{-8}	8.9×10^{-7}	5.8×10^{-8}	2.4×10^{-8}	1.3×10^{-7}
fg	6.8×10^{-11}	1.4×10^{-7}	1.3×10^{-11}	1.7×10^{-13}	1.9×10^{-12}
iter	7	20	10	10	7

TABLE 2.2
B_0 using the identity.

(TP, n, m)	(56, 7, 4)	(78, 5, 3)	(104, 8, 4)	($P1$, 2, 1)	($P2$, 3, 2)
eig (B_0)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
eig (P_0B_0)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
eig (B_k)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
eig (P_kB_k)	(0, 0, 7)	(0, 0, 5)	(0, 0, 8)	(0, 0, 2)	(0, 0, 3)
fh	1.7×10^{-7}	8.4×10^{-10}	4.6×10^{-7}	1.3×10^{-8}	7.4×10^{-11}
fg	7.4×10^{-7}	5×10^{-13}	6.6×10^{-11}	1.1×10^{-10}	2×10^{-12}
iter	12	6	12	7	5

meaning the number of negative, zero, and positive eigenvalues of the corresponding matrix. For instance, if the eigenvalues of H_* are $(2, 1, 0, -1, -3, -4)$, then in the row corresponding to eig (H_*) we will have the triplet (3, 1, 2). An eigenvalue is set to zero if it is less than 10^{-16} in absolute value.

The last four rows in Table 1.1 are used to give more information about the performance of the algorithm:

id distance from x_0 to x_*;
fh the final value of $\|h(x_k)\|$;
fg the final value of $\|g(x_k)\|$;
iter the number of iterations.

The eigenvalues of the matrices $P_k B_k P_k + \nabla g_k \nabla g_k^T$ are given to check the positive definiteness of B_k along the subspace S_k. Let us now comment on the results. The finite difference approximation maintains the same structure on the eigenvalues. Nothing extra was needed for B_0 to be positive definite along S_0. The eigenvalues of B_k and $P_k B_k P_k + \nabla g_k \nabla g_k^T$ for all k, although not given here, were computed and no changes were reported. The eigenvalue structure remained the same during the process, as was expected due to the results of §§ 2 and 3.

Test problem 26 is interesting in itself. It is the only one for which the method failed to converge. The reason is that the Hessian H_* is not positive definite along S_* numerically; that is, one of the eigenvalues of the matrix $P_* H_* P_* + \nabla g_* \nabla g_*^T$ is 0.7×10^{-16}. The finite difference approximation $P_0 B_0 P_0 + \nabla g_0 \nabla g_0^T$ shows this eigenvalue to be 3.8×10^{-13}, and therefore the method diverges.

In test problem 47 (Table 1.1), B_0 is positive definite and so are the remainder of the B_k. However, we still have convergence even though the Hessian H_* that the B_k is approximating is only positive definite on S_*.

Tables 2.1 and 2.2 indicate that the method is quite robust with respect to the initial guess B_0. In fact, we even have convergence for test problem 26. We think this is partly because the matrices B_k remain positive definite throughout the process although H_* is positive definite only on S_*. This is predictable, since $y_k^T w_k > 0$ for all k and B_0 is positive definite.

Acknowledgments. The author thanks J. E. Dennis for the fruitful discussion that led to Theorem 2.3, and Jorge Moré for his helpful comments that led to Lemma 3.2 and made the paper more readable.

REFERENCES

[1] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[2] C. G. BROYDEN, J. E. DENNIS, JR., AND J. J. MORÉ, *On the local and superlinear convergence of quasi-Newton methods*, J. Inst. Math. Appl., 12 (1973), pp. 223-245.

[3] J. E. DENNIS, JR. AND J. J. MORÉ, *A characterization of q-superlinear convergence and its applications to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549-560.

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

[5] R. FONTECILLA, *A Newton-like method for nonlinear constrained optimization*, Tech. Report 1375, Dept. of Computer Science, University of Maryland, College Park, MD, 1984.

[6] ———, *Projected secant methods for nonlinearly constrained optimization*, Tech. Report 1386, Dept. of Computer Science, University of Maryland, College Park, MD, 1984.

[7] A. GRIEWANK AND PH. L. TOINT, *Local convergence analysis for partitioned quasi-Newton updates*, Numer. Math., 39 (1982), pp. 429-448.

[8] W. HOCK AND K. SCHITTKOWSKI, *Test examples for nonlinear programming codes*, Lecture Notes in Econom. and Math. Systems 187, Springer-Verlag, Berlin, 1981.

[9] J. NOCEDAL AND M. OVERTON, *Projected Hessian updating algorithm for nonlinear constrained optimization*, SIAM J. Numer. Anal., 22 (1985), pp. 821-850.

[10] A. STACHURSKI, *Superlinear convergence of Broyden's bounded θ-class of methods*, Math. Programming, 20 (1981), pp. 196-212.

[11] R. A. TAPIA, *Quasi-Newton methods for equality constrained optimization: equivalence of existing methods and a new implementation*, in Nonlinear Programming 3, O. L. Mangasarian, R. Meyer, and S. Robinson, eds., Academic Press, New York, 1978, pp. 125-164.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 53-82, January 1989

005

LOCAL TIME-OPTIMAL FEEDBACK CONTROL OF STRICTLY NORMAL TWO-INPUT LINEAR SYSTEMS*

L. DAVID MEEKER†

Abstract. This paper introduces a new technique for the analysis of time-optimal control of linear systems. A family of easily calculated invariants is developed and, for an important class of two-input systems, is shown to provide a complete description of the time-optimal flow near the origin. The two switching surfaces are described analytically and qualitatively in topological terms. The time-optimal feedback function is defined and analyzed with respect to its complexity and sensitivity to errors in state variable measurement. The results lead to the first explicit construction of a local regular synthesis for multi-input systems of arbitrary order.

Key words. time-optimal feedback, closed-loop, stability

AMS(MOS) subject classifications. primary 49E15, 49E25

1. Introduction. This paper is concerned with the description, synthesis, and classification of time-optimal control systems that can be modeled by equations of the following form:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \tag{1.1a}$$

$$\mathbf{x} \in \mathbb{R}^n, \qquad \mathbf{u} \in \mathscr{C}^r \equiv \{\mathbf{u} \in \mathbb{R}^r : |u_i| \leqq 1,\ i = 1, 2, \cdots, r\}, \tag{1.1b}$$

$$\mathbf{B} = [\mathbf{b}^1, \cdots, \mathbf{b}^r], \quad \mathbf{b}^i \in \mathbb{R}^n, \quad \operatorname{rank}(\mathbf{B}) = r. \tag{1.1c}$$

Time-optimal control has been the subject of considerable research effort during the past 25 years. As a result of these investigations the optimal control functions have been identified and shown to be conceptually simple—piecewise constant with each control component either $+1$ or -1—with switching surfaces $\mathbf{\Omega}_1, \cdots, \mathbf{\Omega}_r$ ($\mathbf{\Omega}_i$ is the set of points of $\mathbb{R}^n$ where the ith control component changes sign) shown to be homeomorphic to a relatively open convex subset of an $(n-1)$-dimensional hyperplane and to divide the **controllable set K** (the set of points of $\mathbb{R}^n$ controllable to $\mathbf{0}$) into two relatively open sets corresponding to the two values $u_i = 1$ and $u_i = -1$ (thus showing that $\mathbf{K} \backslash \mathbf{\Omega}_i$ is the union of two disjoint components) [25].

These results provide an essentially complete qualitative picture of the time-optimal flow (the time-optimal trajectories) in the case of scalar control (when $r = 1$). However, this is far from the case for multidimensional controls ($r > 1$). The presence of two, or more, control dimensions greatly complicates the problem. In this instance several interesting and important questions arise:

(1) How do the switching surfaces interact? What is $\mathbf{\Omega}_i \cap \mathbf{\Omega}_j$ topologically?
(2) Can $\mathbf{\Omega}_i = \mathbf{\Omega}_j$ if $i \neq j$?
(3) What is the topological structure of $\mathbf{\Omega} \equiv \bigcup \mathbf{\Omega}_i$?
(4) How complicated or how simple can the piecewise constant optimal feedback function be? That is, how many "pieces" can there be, or "what is the connectivity of $\mathbf{K} \backslash \mathbf{\Omega}$"?
(5) How sensitive is the time-optimal feedback function to measurement errors or time lags in the feedback loop? In particular, is the time-optimal feedback control system stable with respect to measurement as defined by Hermes [12]?
(6) How are the answers to these questions encoded in the two matrices **A** and **B**?

* Received by the editors September 19, 1979; accepted for publication (in revised form) April 5, 1988.
† Department of Mathematics, University of New Hampshire, Durham, New Hampshire 03824.

In general, the answers to these questions are unknown. With $r>1$ only the case $n=r=2$ is fully understood [3], [4]. Other investigations have been directed toward the local analogues of the questions. That is, procedures for the analysis of the time-optimal flow on $\mathbf{K}(T)$, the ***T*-controllable set** composed of those points of $\mathbb{R}^n$ controllable to $\mathbf{0}$ in time T, in terms of the corresponding relativized switching surfaces $\mathbf{\Omega}_i(T)\equiv\mathbf{\Omega}_i\cap\mathbf{K}(T)$, $i=1,\cdots,r$, and $\mathbf{\Omega}(T)\equiv\mathbf{\Omega}\cap\mathbf{K}(T)$, for sufficiently small T [9], [17]-[19], [21], [22], [26], [27], and [29]. The answers to the local versions of questions (1)-(6) for the minimally controllable systems (a generic class discussed in § 2 herein) appear in [20] and [21] for the case $n=3$, $r=2$, and all but question (5) are answered for the $n=4$, r = 2 case in [22]. The general n, r case is discussed in a preliminary fashion in [19]. In this paper the answers to the local versions of questions (1)-(4) and (6) are answered for the class of strictly normal systems (see [9], [29], and § 2 herein) for systems with arbitrary n and $r=2$. Sufficient conditions for measurement stability (*H*-stability) are also developed for this class of systems.

The analysis and description of the time-optimal flow for these systems is simplified by the introduction of two mathematical structures that as yet have not become familiar in control theory literature. The first of these, the *theory of cell complexes* or, more properly, the *theory of* CW-*complexes* (see [16]) is necessary to describe the collection of points in $\mathbf{K}(T)$ controlled to $\mathbf{0}$ by all control functions having the same specific sequence of control values. Below we show that $\mathbf{K}(T)$ can be decomposed into a collection of well-defined cell complexes each associated with known switching sequences. The description and analysis of the time-optimal flow on each component cell complex are facilitated by the concepts and results of the *theory of semidynamical systems* (see [1]) whose characterization of attracting and invariant sets is critical to the results that follow. The following discussion of the known results for the cases $n=2$ and 3, $r=2$ is intended to illustrate the value of these two mathematical systems for the study and description of time-optimal or, more generally, bang-bang control systems.

In the well-known case where $n=r=2$ the time-optimal flow on $\mathbf{K}(T)$ occurs, for sufficiently small T, in only the two distinct forms shown in Fig. 1. The relative structure of the switching curves $\mathbf{\Omega}_1(T)$ and $\mathbf{\Omega}_2(T)$ is very different in the two cases. For systems of the type shown in Fig. 1(a), such as

$$\mathbf{A}=\begin{bmatrix}0 & -1\\ 1 & 0\end{bmatrix},\qquad \mathbf{B}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

the two curves intersect only at $\mathbf{0}$, while for systems of the type shown in Fig. 1(b), such as

$$\mathbf{A}=\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix},\qquad \mathbf{B}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix},$$

the two curves coincide and $\mathbf{\Omega}_1(T)=\mathbf{\Omega}_2(T)=\mathbf{\Omega}(T)$. $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$ has four components for systems such as those of Fig. 1(a) but only two components for those of Fig. 1(b).

The time-optimal trajectories can be described on the component **ABOA** of Fig. 1(a) by "use control $\mathbf{u}^1$ until you reach curve **BO**, then switch to $\mathbf{u}^2$ and use it until arrival at $\mathbf{0}$" (assuming the proper "naming" of the control vectors). This sequence "$\mathbf{u}^1$ switch to $\mathbf{u}^2$" defines what we will call an optimal control (switching) policy $\mathbf{p}=\langle\mathbf{u}^1,\mathbf{u}^2\rangle$, or, when there is no chance of confusion, simply $\mathbf{p}=\langle 1,2\rangle$. The component **ABOA** is composed of those points of $\mathbf{K}(T)$ controlled to $\mathbf{0}$ by control functions following the policy $\langle 1,2\rangle$ and is, accordingly, denoted by $\mathbf{D}\langle 1,2\rangle$. If we define the other two vertices of the control domain $\mathscr{C}^2$ by $\mathbf{u}^3=-\mathbf{u}^1$ and $\mathbf{u}^4=-\mathbf{u}^2$, the other

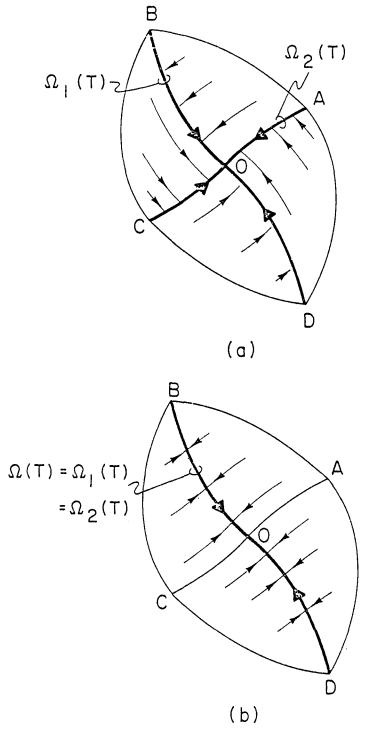

FIG. 1. *The two possible types of time-optimal flow near* $\mathbf{0}$ *for* $n = r = 2$.

components of $\mathbf{K}(T)$ in Fig. 1(a), **BCOB**, **CDOC**, and **DAOD**, are associated with the control policies $\langle 2, 3\rangle$, $\langle 3, 4\rangle$, and $\langle 4, 1\rangle$ and denoted by $\mathbf{D}\langle 2, 3\rangle$, $\mathbf{D}\langle 3, 4\rangle$, and $\mathbf{D}\langle 4, 1\rangle$, respectively. Similarly, $\mathbf{K}(T)$ of Fig. 1(b) can be described as the union of the four components $\mathbf{D}\langle 1, 2\rangle$, $\mathbf{D}\langle 3, 2\rangle$, $\mathbf{D}\langle 3, 4\rangle$, and $\mathbf{D}\langle 1, 4\rangle$.

Each set $\mathbf{D}\langle i, j\rangle$ is homeomorphic to the standard 2-cell $\boldsymbol{\sigma} = \{\mathbf{s} \in \mathbb{R}^2 : 0 \leqq s_1 \leqq s_2 \leqq T\}$ (this is discussed in detail in § 4). For example, $\mathbf{D}\langle 1, 2\rangle$ is the image of $\boldsymbol{\sigma}$ under the map

$$\mathbf{x}(s_1, s_2) = -\left[(\mathbf{u}^1 - \mathbf{u}^2) \int_0^{s_1} e^{-\mathbf{A}\tau} \mathbf{B}\, d\tau + \mathbf{u}^2 \int_0^{s_2} e^{-\mathbf{A}\tau} \mathbf{B}\, d\tau \right] \tag{1.2}$$

and is, therefore, a 2-cell (see [16]) composed of its interior open 2-cell and its three boundary 1-cells—the two constant control trajectories $\mathbf{D}\langle 1\rangle$ and $\mathbf{D}\langle 2\rangle$ (**AO** and **BO**) and the portion of the boundary T-isochrone (**AB**). Thus for each system type, as in Figs. 1(a) and 1(b), $\mathbf{K}(T)$ can be described as a cell complex consisting of the union (actually, in the language of CW-complexes, cellular adjunction) of four 2-cells each associated with a unique time-optimal switching policy (switching sequence). This is an example of the cellular decomposition proved for the general strictly normal system in § 6. However, we shall see that for general n, the maps corresponding to (1.2) of the standard n-cell into $\mathbf{K}(T)$ need not be injective except on the interior of the standard cell—certain boundary cells may coalesce or "lose dimension" under some maps. In such cases it is necessary to consider $\mathbf{K}(T)$ as a CW-complex rather than simply a cell complex.

The difference in the time-optimal flow pattern and the switching curve structure in Figs. 1(a) and 1(b) can be described by noting that the time-optimal trajectories

define a local semidynamical system on $\mathbf{K}(T)\backslash\{\mathbf{0}\}$ (a *semidynamical system* is a generalized dynamical system that permits intersecting trajectories; a *local* semidynamical system is a semidynamical system that has a finite escape time—provided, in this instance, by the optimal response time—for each point of the state space; see [1]).

The switching curve **BO** (or $\mathbf{D}\langle 2\rangle$) of **ABOA** (or $\mathbf{D}\langle 1, 2\rangle$) is the image of the 1-cell $\{(0, s_2): 0 < s_2 \leqq T\}$ of $\boldsymbol{\sigma}$ under (1.2) and is an **attracting set** of the flow while the curve **AO** (or $\mathbf{D}\langle 1\rangle$), the image of the 1-cell $\{(s_2, s_2): 0 < s_2 \leqq T\}$ under (1.2), is an **invariant set** of the flow. Note that $\mathbf{D}\langle 2\rangle$ is an attracting set of $\mathbf{D}\langle 1, 2\rangle$, but an invariant set of the adjoining cell $\mathbf{D}\langle 2, 3\rangle$. Indeed, the difference in the time-optimal flows of the two system types stems precisely from the fact that in systems such as those in Fig. 1(a), *attracting cells always adjoin invariant cells* and vice versa, while in systems such as those in Fig. 1(b), *attracting cells always adjoin attracting cells* and *invariant cells always adjoin invariant cells.* The structure theorem of § 9 extends these results to the general system (1.1) with $r = 2$.

As a consequence of the difference in their optimal flow patterns, the character of the time-optimal feedback function differs substantially between the two system types. In Fig. 1(b) the value it assumes on the switching curves **OB** and **OD** differ from the values assumed on either side of the curves. Obviously, implementation of such a closed-loop system would require infinitely precise measurement of the state variables, since the curves are sets of planar measure zero. On the other hand, the feedback function of systems of the type in Fig. 1(a) are much more tolerant of measurement errors, since small errors in state measurement yield either the correct control or a control value that drives the system across the switching curve and, thereby, increases the probability of determining the proper control. In § 7, where such matters are fully discussed, the feedback function of a system of the type in Fig. 1(a) is said to be **realizable**, while that of a system of the type in Fig. 1(b) is **nonrealizable** on the curve **BOD**.

As is to be expected, the closed-loop time-optimal systems of the two types differ in their response to measurement errors or time-lags in the control loop. As shown in Fig. 2(a) the realizable feedback control leads to a (small) "overshoot" in the trajectory, while in Fig. 2(b) the nonrealizable controller leads to "chatter" or a "sliding" motion back and forth across the time-optimal trajectory.

The system of Fig. 1(a) is actually "stable with respect to measurement" or H-stable in the sense of Hermes [12]. This concept will be discussed in § 11 herein.

Time-optimal control for the case $n = 3$, $r = 2$ is less well known. As shown in [20], strictly normal systems (to be defined below) occur in precisely three canonical and topologically distinct forms. Figure 3 provides sketches of typical switching surface structures (near the target point **0**) for each possible type.

In every case the switching surfaces $\boldsymbol{\Omega}_1(T)$ and $\boldsymbol{\Omega}_2(T)$ are composed of four two-dimensional cells. In Fig. 3(a) the surfaces have no 2-cell in common and intersect only in the 1-cells corresponding to constant control trajectories. This leads to the most complex structure with $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ consisting of six disjoint components (i.e., having connectivity equal to 6). It happens that for such systems the time-optimal feedback function is realizable on $\mathbf{K}(T)$ for sufficiently small T.

In Fig. 3(b), $\boldsymbol{\Omega}_1(T)$ and $\boldsymbol{\Omega}_2(T)$ contain two 2-cells in common. As a result $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ has connectivity equal to 4, and the feedback function is nonrealizable on the two common 2-cells where the closed-loop time-optimal system is "chatter-prone."

In the last case, Fig. 3(c), the two switching surfaces coincide and $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ has connectivity equal to 2. This situation is the three-dimensional analogue of the

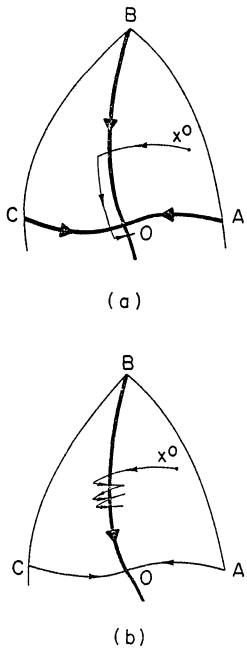

FIG. 2. *Typical responses of the systems of Fig.* 1 *to time lags or errors in state variable measurement in the time-optimal feedback loop.*

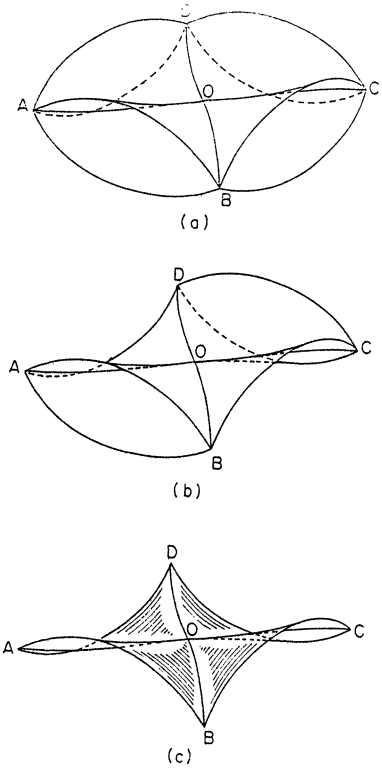

FIG. 3. *Qualitative description of switching surface structures for third-order two-input strictly normal systems.* (a) *The case* $\gamma(1) = \gamma(2) = 1$. (b) *The case* $\gamma(1) = -\gamma(2)$. (c) *The case* $\gamma(1) = \gamma(2) = -1$.

two-dimensional structure of Fig. 1(b) and the closed-loop feedback function is nonrealizable at each point of $\Omega(T)$.

In the sections to follow the analysis that has led to these results is extended to the class of strictly normal nth-order systems having two-dimensional controls. This analysis will show that the structure of the time-optimal flow near the origin is determined completely by geometric relationships among the $2n$ vectors $\mathbf{b}^1$, $\mathbf{Ab}^1$, $\cdots$, $\mathbf{A}^{n-1}\mathbf{b}^1$, $\mathbf{b}^2$, $\mathbf{Ab}^2$, $\cdots$, $\mathbf{A}^{n-1}\mathbf{b}^2$ as expressed by the values of the $n+1$ determinants:

$$\mathbf{d}(j) \equiv \det[\mathbf{b}^1, \mathbf{Ab}^1, \cdots, \mathbf{A}^{j-1}\mathbf{b}^1, \mathbf{b}^2, \mathbf{Ab}^2, \cdots, \mathbf{A}^{n-j-1}\mathbf{b}^2], \qquad j = 0, 1, \cdots, n \tag{1.3}$$

(which, for strictly normal systems, are all nonzero [9], [29]); their signs:

$$\delta(j) \equiv \operatorname{sgn}[\mathbf{d}(j)]; \tag{1.4}$$

and the **structure invariants**:

$$\gamma(j) = \gamma(j; \mathbf{A}, \mathbf{B}) \equiv \delta(j-1) \cdot \delta(j+1), \qquad j = 1, \cdots, n-1, \tag{1.5}$$

$$N(\mathbf{A}, \mathbf{B}) = \sum_{j=1}^{n-1} (1 - \gamma(j; \mathbf{A}, \mathbf{B}))/2, \tag{1.6}$$

$$M(\mathbf{A}, \mathbf{B}) \equiv \sum_{j=1}^{n-1} (1 - \gamma(j; \mathbf{A}, \mathbf{B}))\binom{n-2}{j-1}. \tag{1.7}$$

The vital role these invariants play in the theory is evident from the following theorem describing the local switching surface structure for two-input systems (the concepts of feedback function "realizability" and "Filippov type" are formally defined in § 7).

THEOREM 1.8. *Let system* (1.1) *be strictly normal, let* $r = 2$, *and let* T *be sufficiently small and positive. Then*

(a) $\mathbf{K}(T) \backslash \Omega(T)$ *has connectivity* $2(n - N(\mathbf{A}, \mathbf{B}))$;

(b) $\Omega_1(T) \cap \Omega_2(T)$ *is the union of* $M(\mathbf{A}, \mathbf{B})$ $(n-1)$*-dimensional cells*;

(c) $\Omega_1(T) = \Omega_2(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) = n-1$ (*that is,* $\gamma(1; \mathbf{A}, \mathbf{B}) = \cdots = \gamma(n-1; \mathbf{A}, \mathbf{B}) = -1$);

(d) *The time-optimal feedback function is realizable on* $\mathbf{K}(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) = 0$ (*that is,* $\gamma(1; \mathbf{A}, \mathbf{B}) = \cdots = \gamma(n-1; \mathbf{A}, \mathbf{B}) = 1$);

(e) *The feedback function is of Filippov type on the relative interior of some* $(n-1)$*-dimensional cell of* $\Omega(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) \leqq n-2$.

These results, which are consequences of the cellular decomposition theorem proved in § 6, imply that in the case $n = 2$, only the single invariant $\gamma(1) = \delta(0) \cdot \delta(2)$ is defined and the structures of Figs. 1(a) and 1(b) occur when $\gamma(1)$ equals 1 and -1, respectively. When $n = 3$, two invariants, $\gamma(1) = \delta(0) \cdot \delta(2)$ and $\gamma(2) = \delta(1) \cdot \delta(3)$, are defined. The switching surfaces have the structure of Fig. 3(a) when $\gamma(1) = \gamma(2) = 1$, the structure of Fig. 3(b) when $\gamma(1) = -\gamma(2)$, and the structure of Fig. 3(c) when $\gamma(1) = \gamma(2) = -1$.

While complex approximations and combinatorial analysis of §§ 3–6 underlie the proof of Theorem 1.8, the final result is a powerful and simple tool for the design and analysis of time-optimal control systems. As the following examples illustrate, the

theorem provides a series of easily calculated invariants which provide a complete qualitative description of the local switching surface structure and of the time-optimal feedback function.

Example A. The system

$$\mathbf{A}=\begin{bmatrix}0 & 0 & 1\\ 1 & 0 & -2\\ 0 & 1 & 0\end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix}1 & 0\\ 0 & 0\\ 0 & 1\end{bmatrix}$$

has $\mathbf{d}(0)=\mathbf{d}(1)=\mathbf{d}(3)=1$, $\mathbf{d}(2)=2$, and $\boldsymbol{\gamma}(1)=\boldsymbol{\gamma}(2)=1$, which implies the switching surface structure of Fig. 3(a) with $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ having maximal connectivity of 6 and $\boldsymbol{\Omega}_1(T)\cap\boldsymbol{\Omega}_2(T)$ consisting only of 1-cells. In addition, as will be shown in § 11, this system is locally measurement stable.

Example B. The system

$$\mathbf{A}=\begin{bmatrix}0 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 0\end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix}1 & 0\\ 0 & 0\\ 0 & 1\end{bmatrix}$$

has $\mathbf{d}(0)=\mathbf{d}(1)=1$, $\mathbf{d}(2)=\mathbf{d}(3)=-1$, and $\boldsymbol{\gamma}(1)=\boldsymbol{\gamma}(2)=-1$. This configuration implies the switching surface structure of Fig. 3(c) with $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ having minimal connectivity of 2 and $\boldsymbol{\Omega}_1(T)=\boldsymbol{\Omega}_2(T)$. In this situation the closed-loop system is "chatter-prone" (see Fig. 2(b)) at each point of $\boldsymbol{\Omega}(T)$.

Example C. The small angle roll-yaw motions of a satellite in circular orbit about a spherical earth in normalized time are of the form (1.1) with

$$\mathbf{A}=\begin{bmatrix}0 & 1 & 0 & 0\\ -\alpha_1 & 0 & 0 & 1-\alpha_1\\ 0 & 0 & 0 & 1\\ 0 & \alpha_2-1 & -4\alpha_2 & 0\end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 0\\ 0 & 1\end{bmatrix},$$

with the two parameters α_1 and α_2 constrained by the conditions $0<\alpha_2<\alpha_1<1$ [22]. For this fourth-order system $\mathbf{d}(4)=4\alpha_2(1-\alpha_2)^2$, $\mathbf{d}(3)=1-\alpha_2$, $\mathbf{d}(2)=1$, $\mathbf{d}(1)=1-\alpha_1$, and $\mathbf{d}(0)=\alpha_1(1-\alpha_1)^2$ so, under the constraints above, $\boldsymbol{\gamma}(1)=\boldsymbol{\gamma}(2)=\boldsymbol{\gamma}(3)=1$. Thus, by Theorem 1.8, since $N(\mathbf{A},\mathbf{B})=M(\mathbf{A},\mathbf{B})=0$, $\mathbf{K}(T)\backslash\boldsymbol{\Omega}(T)$ has connectivity equal to 8, its (three-dimensional) switching surfaces intersect only in cells of dimensions 1 and 2, and its time-optimal feedback function is realizable on all of $\mathbf{K}(T)$ (i.e., Fig. 2(a)) and, therefore, tolerant of errors in the feedback loop.

2. Strictly normal and minimally controllable systems. The method of analysis described in this paper is applicable to a generic class of time-optimal control systems which have been termed "minimally controllable" [21].

DEFINITION 2.1. System (1.1) is said to be minimally controllable if there exists a neighborhood $\mathbf{U}$ of $\mathbf{0}\in\mathbb{R}^n$ such that for each $\mathbf{x}\in\mathbf{U}$ there exists an extremal control function $\mathbf{v}(\cdot,\mathbf{x})$ that controls $\mathbf{x}$ to the origin in minimum time and has no more than $n-1$ switches.

Remark 2.2. Since a system is controllable if $\mathbf{K}(T)$ contains a neighborhood of $\mathbf{0}$ [14], minimally controllable systems are controllable, but need not be normal (i.e., have unique time-optimal control functions [14]). For instance, the linearized rendezvous problem [18] presents an example of a minimally controllable but nonnormal system.

Remark 2.3. In the scalar control case ($r=1$) or the case $n=2$, a system is minimally controllable if and only if it is controllable. Sufficient conditions for minimal controllability have appeared in [17], [19], and [20].

Hájek [9] has called the general system (1.1) with r-dimensional controls **strictly normal** if, for any collection of integers, $j_1 \geqq 0, \cdots, j_r \geqq 0$ satisfying $j_1 + \cdots + j_r = n$, the n vectors $\mathbf{A}^m\mathbf{b}^i$, $0 \leqq m \leqq j_{i-1}$, $i = 1, 2, \cdots, r$, are linearly independent. For example, when $r=2$, the case of interest here, a system is strictly normal if and only if all the determinants $\mathbf{d}(0), \mathbf{d}(1), \cdots, \mathbf{d}(n)$ of (1.3) are nonzero.

Yeung [29] has shown that normal systems are strictly normal if and only if they are minimally controllable. He has also shown that for such systems and for small T, $\mathbf{K}(T)$ is the union of $2r^{n-1}$ sets, which Hájek [9] called "terminal manifolds," each of which has a nonempty interior and is associated with a distinct control-switching sequence. In the following sections these control sequences are identified for the case $r=2$ and the corresponding 2^n terminal manifolds shown to form natural groups whose interiors form the components of $\mathbf{K}(T)\backslash\mathfrak{Q}(T)$.

3. Identification of time-optimal control functions. In this section we identify, for strictly normal two-input systems, all time-optimal control functions having small response times. These controls are shown to be naturally grouped: first, by the number of first coordinate switches and, second, by the actual sequence of control values assumed by the functions.

Our primary tool in this investigation is, of course, the minimum principle of Pontryagin [14] which, in this instance, implies that a control function $\mathbf{v}$ is extremal for a point $\mathbf{x}$ (i.e., controls $\mathbf{x}$ to $\mathbf{0}$ in the shortest possible time) if and only if it satisfies

$$\boldsymbol{\lambda}' e^{-\mathbf{A}t}\mathbf{B}\mathbf{v}(t) = \min\{\boldsymbol{\lambda}' e^{-\mathbf{A}t}\mathbf{B}\mathbf{u}: \mathbf{u} \in \mathscr{C}^2\} \tag{3.1}$$

almost everywhere on $0 \leqq t \leqq T$, where T is the optimal response time for $\mathbf{x}$. The vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ is an outernormal of a supporting hyperplane of $\mathbf{K}(T)$ at $\mathbf{x}$ [14]. Such a vector $\boldsymbol{\lambda}$ will be said to *generate control functions* $\mathbf{v}$ *and* $-\mathbf{v}$ *over the interval* [0, T] if $\boldsymbol{\lambda}$ and $\mathbf{v}$ satisfy (3.1). (It should be noted that, with this definition, $-\boldsymbol{\lambda}$ also generates $\mathbf{v}$ and $-\mathbf{v}$ over $[0, T]$.)

If $\boldsymbol{\lambda}$ *generates* $\mathbf{v} = (v_1, v_2)$ (we identify points in $\mathbb{R}^m$ with $m \times 1$ vectors) then (3.1) implies

$$v_i(t) = -\operatorname{sgn}[\boldsymbol{\lambda}' e^{-\mathbf{A}t}\mathbf{b}^i], \qquad i = 1, 2, \tag{3.2}$$

whenever $\boldsymbol{\lambda}' e^{-\mathbf{A}t}\mathbf{b}^i \neq 0$. Since the functions $t \mapsto \boldsymbol{\lambda}' e^{-\mathbf{A}t}\mathbf{b}^i$, $i=1, 2$, are analytic, they either vanish identically or have isolated zeros. In the latter instance such zeros, where v_1 or v_2 may change sign, correspond to possible *switching points* of $\mathbf{v}$. If neither of the two functions vanishes identically, $\mathbf{v}$ is uniquely defined by (3.2) and the additional requirement of continuity from the right. *The assumption of right continuity will be implicit below.*

Because of (3.1) each extremal control function is generated by some vector on the unit sphere in $\mathbb{R}^n$. Moroz [25] and others have studied the set of extremal functions using parametrizations of the sphere. Unfortunately, while these global parametrizations yield a description of the switching behavior of the control functions, they do so in an obscure and complicated fashion.

An alternate method of parametrization is developed here. While it is not global in scope—n parametric families are required to describe the extremal control functions on $[0, T]$ for small T—it has the virtue of being intimately related to the switching behavior of the control functions themselves. This description is based on the fact that the control function generated by the vector $\boldsymbol{\lambda}$ switches in component i ($i=1$ or 2) at

time s if and only if $\boldsymbol{\lambda}$ is orthogonal to the vector

$$\mathbf{E}^i(s) \equiv e^{-\mathbf{A}s}\mathbf{b}^i.$$

With this observation it is clear that $\mathbf{v}$ (in (3.1)) switches in components $\alpha, \beta, \gamma, \cdots$ (where $\alpha, \beta, \gamma, \cdots = 1$ or 2) at times $s_1, s_2, s_3, \cdots$ only if $\boldsymbol{\lambda}$ is orthogonal to $\mathbf{E}^\alpha(s_1), \mathbf{E}^\beta(s_2), \mathbf{E}^\gamma(s_3), \cdots$. The notation of exterior algebra [7] is useful in expressing such relationships. For example, if $\mathbf{x}^1, \cdots, \mathbf{x}^{n-1}$ are linearly independent vectors, the $(n-1)$-vector $\mathbf{x}^1 \wedge \cdots \wedge \mathbf{x}^{n-1}$ may be identified with the unique vector $\boldsymbol{\lambda} \in \mathbb{R}^n$ that satisfies

$$\boldsymbol{\lambda}'\mathbf{x} = \mathbf{x}^1 \wedge \cdots \wedge \mathbf{x}^{n-1} \wedge \mathbf{x} = \det[\mathbf{x}^1, \cdots, \mathbf{x}^{n-1}, \mathbf{x}],$$

for all $\mathbf{x}$ in $\mathbb{R}^n$. Clearly $\boldsymbol{\lambda}$ is orthogonal to $\mathbf{x}^1, \cdots, \mathbf{x}^{n-1}$ and $\boldsymbol{\lambda}'\mathbf{x} = 0$ only for those $\mathbf{x}$ in the span of $\mathbf{x}^1, \cdots, \mathbf{x}^{n-1}$.

The relationships among the outernormals of $\mathbf{K}(T)$ (the $\boldsymbol{\lambda}$'s), the switching times s_j, and the vectors $\mathbf{E}^i(s_j)$ permit (in the case of strict normality) the identification of all the time-optimal control functions over a sufficiently small interval $[0, T]$. For example, in the three-dimensional case, any $\boldsymbol{\lambda}$ generating a control function having a switch on the first coordinate when $t = s_1$, and a switch on the second coordinate when $t = s_2$, *must* be orthogonal to $\mathbf{E}^1(s_1)$ and $\mathbf{E}^2(s_2)$ and thus *must* be a multiple of $\boldsymbol{\lambda}^1(s_1, s_2) \equiv \mathbf{E}^1(s_1) \wedge \mathbf{E}^2(s_2)$.

The control function $\mathbf{v} = (v_1, v_2)$ generated by $\boldsymbol{\lambda}^1(s_1, s_2)$ is determined (via (3.2)) by the signs of

$$\begin{aligned}
-\boldsymbol{\lambda}^1(s_1, s_2)'\mathbf{E}^1(t) &= -(\mathbf{b}^1 - \mathbf{A}\mathbf{b}^1 s_1 + \cdots) \wedge (\mathbf{b}^2 - \mathbf{A}\mathbf{b}^2 s_2 + \cdots) \wedge (\mathbf{b}^1 - \mathbf{A}\mathbf{b}^1 t + \cdots) \\
&= (s_1 - t)[\mathbf{b}^1 \wedge \mathbf{A}\mathbf{b}^1 \wedge \mathbf{b}^2 + \cdots] = (s_1 - t)[\det[\mathbf{b}^1, \mathbf{A}\mathbf{b}^1, \mathbf{b}^2] + \cdots] \\
&= (s_1 - t)[\mathbf{d}(2) + \cdots]
\end{aligned}$$

and

$$\begin{aligned}
-\boldsymbol{\lambda}^1(s_1, s_2)'\mathbf{E}^2(t) &= -(\mathbf{b}^1 - \mathbf{A}\mathbf{b}^1 s_1 + \cdots) \wedge (\mathbf{b}^2 - \mathbf{A}\mathbf{b}^2 s_2 + \cdots) \wedge (\mathbf{b}^2 - \mathbf{A}\mathbf{b}^2 t + \cdots) \\
&= (s_2 - t)[-\mathbf{b}^1 \wedge \mathbf{b}^2 \wedge \mathbf{A}\mathbf{b}^2 + \cdots] = (s_2 - t)[-\mathbf{d}(1) + \cdots].
\end{aligned}$$

Therefore, the optimal control functions generated by $\boldsymbol{\lambda}^1(s_1, s_2)$ (which, we recall, are the only functions with a single switch on each coordinate) switch when $t = s_1$ and s_2 and, over a sufficiently small interval, $0 \leqq t \leqq T$, describe the control sequences (recall $\boldsymbol{\delta}(i) = \operatorname{sgn}[\mathbf{d}(i)]$)

$$(-\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)) \to (\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)) \to (\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)),$$

or

$$(\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)) \to (-\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)) \to (-\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)),$$

if $0 < s_1 < s_2 < T$, and the sequences

$$(-\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)) \to (-\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)) \to (\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)),$$

or

$$(\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1)) \to (\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)) \to (-\boldsymbol{\delta}(2), \boldsymbol{\delta}(1)),$$

if $0 < s_2 < s_1 < T$. Thus, of the eight possible control sequences having a switch in each coordinate, only these four are time-optimal. Note that the controls always begin with $\pm(-\boldsymbol{\delta}(2), \boldsymbol{\delta}(1))$ and end with $\pm(\boldsymbol{\delta}(2), -\boldsymbol{\delta}(1))$. Obviously control functions with other switching patterns (e.g., two switches on coordinate 1) can be identified in the same way. Their sequences of control values will be completely determined by

$\delta(1), \delta(2), \delta(0) = \operatorname{sgn}[\mathbf{b}^2 \wedge \mathbf{A}\mathbf{b}^2 \wedge \mathbf{A}^2\mathbf{b}^2]$ and $\delta(3) = \operatorname{sgn}[\mathbf{b}^1 \wedge \mathbf{A}\mathbf{b}^1 \wedge \mathbf{A}^2\mathbf{b}^1]$. This analysis underlies the results of [17], [20], and [21]. The same approach will now be used to identify optimal control switching sequences for the general nth-order two-input strictly normal system.

Any control function that switches in the first coordinate at times $s_1, \cdots, s_j$ and in the second coordinate at times $s_{j+1}, \cdots, s_{n-1}$ must be generated by a multiple of the vector

$$\boldsymbol{\lambda}(\mathbf{s}; j) \equiv \mathbf{E}^1(s_1) \wedge \cdots \wedge \mathbf{E}^1(s_j) \wedge \mathbf{E}^2(s_{j+1}) \wedge \cdots \wedge \mathbf{E}^2(s_{n-1}), \tag{3.3}$$

for $\mathbf{s} = (s_1, \cdots, s_n) \in \mathbb{R}^n$.

The sequence of control values assumed by a control function generated by $\boldsymbol{\lambda}(\mathbf{s}, j)$ is determined by the signs of $\boldsymbol{\lambda}(\mathbf{s}, j)'\mathbf{E}^i(t)$, $i = 1, 2$, for small values of t. From (3.3) it can be seen that these functions depend on k-vector factors of the form

$$\bigwedge_{j=1}^{k} \mathbf{E}^i(a_j) = \bigwedge_{j=1}^{k} \left(\sum_{m=0}^{\infty} (-1)^m \frac{\mathbf{A}^m \mathbf{b}^i a_j^m}{m!} \right). \tag{3.4}$$

This function is clearly analytic (even entire) in each a_j and vanishes when any two a's are equal. It therefore has the Vandermonde function

$$\operatorname{vdm}(a_1, \cdots, a_k; k) \equiv \prod_{j=1}^{k} \prod_{m=j+1}^{k} (a_m - a_j)$$

as a factor. The leading k-vector term in the expansion is easily seen to be $\mathbf{b}^i \wedge \mathbf{A}\mathbf{b}^i \wedge \mathbf{A}^2\mathbf{b}^i \wedge \cdots \wedge \mathbf{A}^{k-1}\mathbf{b}^i$.

With these two facts in mind it is easy to see that

$$\bigwedge_{j=1}^{k} \mathbf{E}^i(a_j) = \frac{(-1)^{k(k-1)/2}}{1! \cdot 2! \cdot \cdots \cdot (k-1)!} \operatorname{vdm}(a_1, \cdots, a_k)(\mathbf{b}^i \wedge \mathbf{A}\mathbf{b}^i \wedge \cdots \wedge \mathbf{A}^{k-1}\mathbf{b}^i + \cdots). \tag{3.5}$$

Now, in order to determine the sign sequences of $\boldsymbol{\lambda}(\mathbf{s}, j)'\mathbf{E}^1(t)$ and $\boldsymbol{\lambda}(\mathbf{s}, j)'\mathbf{E}^2(t)$, which determine the extremal controls generated by $\boldsymbol{\lambda}(\mathbf{s}, j)$, we assume that the switching times satisfy

$$0 < s_1 < \cdots < s_j < s_n < T \quad \text{and} \quad 0 < s_{j+1} < \cdots < s_{n-1} < s_n < T. \tag{3.6}$$

Then, using (3.5), we have

$$\begin{aligned} \boldsymbol{\lambda}(\mathbf{s}, j)'\mathbf{E}^1(t) &= (-1)^{n-1}\mathbf{E}^1(t) \wedge \bigwedge_{m=1}^{j} \mathbf{E}^1(s_m) \wedge \bigwedge_{m=j+1}^{n-1} \mathbf{E}^2(s_m) \\ &= (-1)^p \cdot \alpha \cdot \operatorname{vdm}(t, s_1, \cdots, s_j; j+1) \cdot \operatorname{vdm}(s_{j+1}, \cdots, s_{n-1}; n-j-1) \\ &\qquad \cdot [\mathbf{b}^1 \wedge \cdots \wedge \mathbf{A}^j\mathbf{b}^1 \wedge \mathbf{b}^2 \wedge \cdots \wedge \mathbf{A}^{n-j-2}\mathbf{b}^2 + \mathbf{O}(T)] \\ &= (-1)^p \cdot \alpha \cdot V(\mathbf{s}; j) \cdot \prod_{m=1}^{j} (s_m - t) \cdot [\mathbf{d}(j+1) + \mathbf{O}(T)] \end{aligned} \tag{3.7}$$

where

$$p = n - 1 + j(j+1)/2 + (n-j-1)(n-j-2)/2,$$

$$\alpha = 1/[1! \cdot 2! \cdot \cdots \cdot j! \cdot 1! \cdot 2! \cdot \cdots \cdot (n-j-2)!],$$

$$V(\mathbf{s}; j) = \operatorname{vdm}(s_1, \cdots, s_j; j) \cdot \operatorname{vdm}(s_{j+1}, \cdots, s_{n-1}; n-j-1).$$

Similarly, we find

$$\begin{aligned}\boldsymbol{\lambda}(\mathbf{s},j)'\mathbf{E}^2(t) &= (-1)^{n-j-1}\bigwedge_{m=1}^{j}\mathbf{E}^1(s_m)\wedge\mathbf{E}^2(t)\wedge\bigwedge_{m=j+1}^{n-1}\mathbf{E}^2(s_m)\\ &= (-1)^q\cdot\beta\cdot V(\mathbf{s},j)\cdot\prod_{m=j+1}^{n-1}(s_m-t)\cdot[\mathbf{d}(j)+\mathbf{O}(T)]\end{aligned}\tag{3.8}$$

where

$$q=n-j-1+j(j-1)/2+(n-j)(n-j-1)/2=-2j+p+n-j-1,$$
$$\beta=\alpha\cdot j!/(n-j-1)!.$$

From (3.7) and (3.8) we see, after canceling common sign factors, that if $\mathbf{d}(j)$ and $\mathbf{d}(j+1)$ are both nonzero and T is sufficiently small (so that $\operatorname{sgn}[\mathbf{d}(j)+\mathbf{O}(T)]=\boldsymbol{\delta}(j)$ and $\operatorname{sgn}[\mathbf{d}(j+1)+\mathbf{O}(T)]=\boldsymbol{\delta}(j+1)$), $\boldsymbol{\lambda}(\mathbf{s},j)$ generates a unique pair of control functions over the interval $0\leqq t\leqq s_n\leqq T$ which begin with the controls

$$\pm\mathbf{u}(1,j)\equiv\pm(\boldsymbol{\delta}(j+1),(-1)^{n-j-1}\boldsymbol{\delta}(j))\tag{3.9a}$$

and, after j changes of sign on the first coordinate (at $t=s_1,\cdots,s_j$) and $n-j-1$ changes on the second coordinate (at $t=s_{j+1},\cdots,s_{n-1}$), terminate with the controls

$$\pm\mathbf{u}(n,j)\equiv\pm((-1)^j\boldsymbol{\delta}(j+1),\boldsymbol{\delta}(j)).\tag{3.9b}$$

The requirement (3.6) is not, in fact, necessary. Indeed, (3.5), (3.7), and (3.8) show that

$$\boldsymbol{\lambda}^*(\mathbf{s},j)\equiv\boldsymbol{\lambda}(\mathbf{s},j)/[\operatorname{vdm}(s_1,\cdots,s_j;j)\cdot\operatorname{vdm}(s_{j+1},\cdots,s_{n-1};n-j-1)]\tag{3.10}$$

has removable singularities and is well defined at each point of the boundary of the n-dimensional cell complex

$$\boldsymbol{\Delta}(j,T)\equiv\{\mathbf{s}\in\mathbb{R}^n:0\leqq s_1\leqq\cdots\leqq s_j\leqq s_n\leqq T,0\leqq s_{j+1}\leqq\cdots\leqq s_n\leqq T\}\tag{3.11}$$

and generates the same control functions as $\boldsymbol{\lambda}(\mathbf{s},j)$ for points of its interior (that satisfy (3.6)).

These results are summarized in the following lemma.

LEMMA 3.12. *If*

(a) *The two determinants* $\mathbf{d}(j)$ *and* $\mathbf{d}(j+1)$ (*see* (1.3)) *are both nonzero and have signs* $\boldsymbol{\delta}(j)$ *and* $\boldsymbol{\delta}(j+1)$;

(b) T *is sufficiently small*;

(c) $\mathbf{s}\in\boldsymbol{\Delta}(j,T)=\{\mathbf{s}\in\mathbb{R}^n:0\leqq s_1\leqq\cdots\leqq s_j\leqq s_n\leqq T,0\leqq s_{j+1}\leqq\cdots\leqq s_n\leqq T\}$,

then the vector $\boldsymbol{\lambda}^*(\mathbf{s},j)$ (*see* (3.10)) *generates two time-optimal control functions,* $\mathbf{v}(\cdot;\mathbf{s},j)$ *and* $-\mathbf{v}(\cdot;\mathbf{s},j)$, *over the interval* $0\leqq t\leqq s_n$. *For those points* $\mathbf{s}$ *in the interior of* $\boldsymbol{\Delta}(j,T)$ *these functions satisfy the following*:

(1) $\mathbf{v}(t;\mathbf{s},j)=\mathbf{u}(1,j)\equiv(\boldsymbol{\delta}(j+1),(-1)^{n-j-1}\boldsymbol{\delta}(j))$, $0<t<\min(s_1,s_{j+1})$;

(2) $\mathbf{v}(\cdot;\mathbf{s},j)$ *changes sign in the first coordinate when* $t=s_1,s_2,\cdots,s_j$ *and in the second coordinate when* $t=s_{j+1},s_{j+2},\cdots,s_{n-1}$;

(3) $\mathbf{v}(t;\mathbf{s},j)=\mathbf{u}(n,j)\equiv((-1)^j\boldsymbol{\delta}(j+1),\boldsymbol{\delta}(j))$ *when* $\max(s_j,s_{n-1})<t<s_n$.

By Lemma 3.12, each point of the interior, $\boldsymbol{\Delta}(j,T)^0$, of $\boldsymbol{\Delta}(j,T)$ corresponds to a unique pair of time-optimal control functions satisfying the conclusions (1)-(3). If $\mathbf{s}\in\boldsymbol{\Delta}(j,T)^0$ has distinct coordinates, these control functions have $n-1$ switches and take on n values from among the extreme points of $\mathscr{C}^2$. In this case a typical control sequence for $\mathbf{v}(\cdot;\mathbf{s},j)$ is of the form $\mathbf{u}(1,j),\mathbf{u}^2,\cdots,\mathbf{u}^{n-1},\mathbf{u}(n,j)$ with the values of $\mathbf{u}^2,\cdots,\mathbf{u}^{n-1}$ determined by the relative order of $s_1,\cdots,s_j$ and $s_{j+1},\cdots,s_{n-1}$. For example, if $s_1<s_{j+1}$, $\mathbf{u}^2=(-\boldsymbol{\delta}(j+1),(-1)^{n-j-1}\boldsymbol{\delta}(j))$; while if $s_{j+1}<s_1$, $\mathbf{u}^2=(\boldsymbol{\delta}(j+1),(-1)^{n-j}\boldsymbol{\delta}(j))$. If some $s_\alpha=s_\beta$, $1\leqq\alpha\leqq j<\beta\leqq n-1$, then switches will occur

simultaneously on both coordinates, and one or more of the $\mathbf{u}^i$'s will be missing from the sequence of control values.

To be more precise, suppose that $\mathbf{s} \in \Delta(j, T)^0$ has distinct coordinates, and let $s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n-1)}$ be the first $n-1$ coordinates of $\mathbf{s}$ arranged in order of increasing magnitude. (Note that $s_{\pi(1)}$ must equal the smaller of s_1 and s_{j+1} while $s_{\pi(n-1)}$ must equal the larger of s_j and s_{n-1}.) The control function $\mathbf{v}(\cdot; \mathbf{s}, j)$ is then given by

$$
\begin{aligned}
\mathbf{v}(t; \mathbf{s}, j) &= \mathbf{u}(1, j), && 0 \leqq t < s_{\pi(1)}, \\
&= \mathbf{u}^2, && s_{\pi(1)} \leqq t < s_{\pi(2)}, \\
&\vdots && \vdots \\
&= \mathbf{u}^i, && s_{\pi(i-1)} \leqq t < s_{\pi(i)}, \\
&\vdots && \vdots \\
&= \mathbf{u}^n, && s_{\pi(n-1)} \leqq t \leqq s_n.
\end{aligned} \tag{3.13}
$$

The process of ordering the coordinates of those $\mathbf{s}$ in $\Delta(j, T)^0$ whose coordinates are distinct defines permutations of $\{1, 2, \cdots, n-1\}$ that are characterized by the fact that their inverses are order-preserving on the subsets $\{1, 2, \cdots, j\}$ and $\{j+1, \cdots, n-1\}$. There are precisely $\binom{n-1}{j} = (n-1)!/(j!(n-1-j)!)$ such ***j*-permissible permutations.** From (3.13), it is clear that if $\mathbf{s}^1$ and $\mathbf{s}^2$ are two different points of $\Delta(j, T)^0$ which define the same j-permissible permutation, π, then the relative order of all coordinates s_α and s_β, with $\alpha \leqq j < \beta$, are the same, and thus the corresponding control functions $\mathbf{v}(\cdot, \mathbf{s}^1, j)$ and $\mathbf{v}(\cdot, \mathbf{s}^2, j)$ assume the same sequence of control values but differ in at least one switching time. Let

$$
\Delta(j, T; \pi)^0 = \{\mathbf{s} \in \Delta(j, T)^0 \colon s_\alpha \neq s_\beta,\ 1 \leqq \alpha, \beta \leqq n-1, \text{ and } \pi \text{ orders } s_1, \cdots, s_{n-1}\} \tag{3.14}
$$

and let $\Delta(j, T; \pi)$ be its closure. The mapping

$$
\mathbf{s} \mapsto (s_{\pi(1)}, \cdots, s_{\pi(n-1)}, s_n)
$$

is a cellular homeomorphism [16] (under their natural topologies) of $\Delta(j, T; \pi)^0$ ($\Delta(j, T; \pi)$) and the open n-cell

$$
\sigma(n, T)^0 = \{\mathbf{y} \in \mathbb{R}^n \colon 0 < y_1 < \cdots < y_n < T\}
$$

(closed n-cell $\sigma(n, T)$).

In this way the interior of $\Delta(j, T)$ is seen to be the disjoint union of the $\binom{n-1}{j}$ open cells, $\{\Delta(j, T; \pi)^0 \colon \pi\ j\text{-permissible}\}$, and their common boundary $(n-1)$-cells defined by equalities of the form $s_\alpha = s_\beta$, where $1 \leqq \alpha \leqq j < \beta \leqq n-1$. The time-optimal controls, $\mathbf{v}(\cdot, \mathbf{s}, j)$, assume, for all $\mathbf{s} \in \Delta(j, T; \pi)^0$, the same sequence of n control values, $\mathbf{p}(\pi, j) \equiv \langle \mathbf{u}(1, j), \mathbf{u}^2, \cdots, \mathbf{u}^{n-1}, \mathbf{u}(n, j) \rangle$. Such sequences need to be formally identified if we are to understand the time-optimal flow.

DEFINITION 3.15. A sequence of control values, $\langle \mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^k \rangle$, which describes the sequential values assumed (each on a set of positive measure) by a time-optimal control function $\mathbf{v}$, and which satisfies the additional "policy condition" that $\mathbf{u}^i \neq \mathbf{u}^{i+1}$, $i = 1, \cdots, k-1$ (i.e., that it be such a sequence having minimal length), is called the (time-optimal) **switching policy of order k** or **k-policy** (or, simply, **switching policy**) of $\mathbf{v}$.

With this notation, for each $\mathbf{s}$ in $\Delta(j, T; \pi)^0$, $\mathbf{v}(\cdot; \mathbf{s}, j)$ has the optimal n-policy, $\mathbf{p}(j, \pi)$, and $-\mathbf{v}(\cdot; \mathbf{s}, j)$ the optimal n-policy $-\mathbf{p}(j, \pi) = \langle -\mathbf{u}(1, j), \cdots, -\mathbf{u}(n, j) \rangle$. When $\mathbf{s}$ lies in the boundary of $\Delta(j, T; \pi)$ (where one or more equalities of the form $s_{\pi(1)} = 0$, $s_{\pi(i)} = s_{\pi(i+1)}$, $s_{\pi(n-1)} = s_n$ hold), examination of (3.7) and (3.8) shows that the time-optimal control functions $\pm\mathbf{v}(\cdot; \mathbf{s}, j)$, generated by $\boldsymbol{\lambda}^*(\mathbf{s}, j)$ are also described by (3.13) and have switching policies that are **subpolicies** of $\pm\mathbf{p}(j, \pi)$ in the obvious sense.

Of particular importance are the subpolicies of order $n-1$ of the n-policies $\pm\mathbf{p}(j,\pi)$, of which there are always at least two (unless $n=1$). Namely, in the case of $\mathbf{p}(j,\pi)$, $\mathbf{p}(j,\pi;\mathrm{A})\equiv\langle\mathbf{u}^2,\cdots,\mathbf{u}(n,j)\rangle$ and $\mathbf{p}(j,\pi;\mathrm{I})\equiv\langle\mathbf{u}(1,j),\mathbf{u}^2,\cdots,\mathbf{u}^{n-1}\rangle$, where "A" and "I" indicate the attracting (A) and invariant (I) character of the associated cells in $\mathbf{K}(T)$ to be defined and discussed in §§ 5 and 6.

These two subpolicies may be the only ones of order $n-1$. For example, suppose $n=4$ and $j=0$. Then, for $\mathbf{s}$ in the interior of $\Delta(0,T)=\{\mathbf{s}\colon 0\leqq s_1\leqq s_2\leqq s_3\leqq s_4\leqq T\}$, $\boldsymbol{\lambda}(\mathbf{s},0)$ will generate controls that switch only in the second coordinate and $\mathbf{v}(\cdot\,;\mathbf{s},0)$ will have a switching policy of the form $\mathbf{p}=\langle\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^1,\mathbf{u}^2\rangle$ with $\mathbf{u}^1=\mathbf{u}(1,0)=(\boldsymbol{\delta}(1),-\boldsymbol{\delta}(0))$ and $\mathbf{u}^2=(\boldsymbol{\delta}(1),\boldsymbol{\delta}(0))$. Since the identity is the only zero permissible permutation, $\mathbf{p}$ and $-\mathbf{p}$ are the only time-optimal switching policies of order 4 when $j=0$.

The policy $\mathbf{p}$ has the two subpolicies of order 3: $\mathbf{p}(\mathrm{A})=\langle\mathbf{u}^2,\mathbf{u}^1,\mathbf{u}^2\rangle$, associated with the 3-cell of $\Delta(0,T)$ having $s_1=0$, and $\mathbf{p}(\mathrm{I})=\langle\mathbf{u}^1,\mathbf{u}^2,\mathbf{u}^1\rangle$, associated with the 3-cell defined by $s_3=s_4$. Since a subpolicy must satisfy the "policy condition" of (3.15), these are the only subpolicies of order 3. Points of other boundary cells of $\Delta(0,T)$ generate, through $\boldsymbol{\lambda}^*(\cdot\,,0)$, controls with policies of order 1 or 2. For example, the policy associated with the 3-cell of $\Delta(0,T)$ defined by $s_1=s_2$ and the 2-cell defined by $s_1=s_2=s_3$ is the order-2 policy, $\langle\mathbf{u}^1,\mathbf{u}^2\rangle$. Clearly, the controls generated by $\boldsymbol{\lambda}^*(\cdot\,,0)$ (or $\boldsymbol{\lambda}^*(\cdot\,,n-1)$) will exhibit similar behavior for any value of $n>2$.

The results derived above concerning the time-optimal control functions for strictly normal systems are summarized in the following proposition.

PROPOSITION 3.16. *Let $r=2$ and system (1.1) be strictly normal so that the determinants $\mathbf{d}(j)\neq 0$, $j=0,1,\cdots,n$. Then there exists a $T>0$ such that, for each $j=0,1,\cdots,n-1$, the vectors $\boldsymbol{\lambda}^*(\mathbf{s},j)$, $\mathbf{s}\in\Delta(j,T)$ (see (3.3) and (3.10)) generate time-optimal control functions, $\pm\mathbf{v}(\cdot\,;\mathbf{s},j)$, over the interval $[0,T]$ with the following properties:*

(1) *If $\mathbf{s}$ is in the interior of $\Delta(j,T)$, $\mathbf{v}(\cdot\,;\mathbf{s},j)$ is unique and if, in addition, $\mathbf{s}$ has distinct coordinates, then $\mathbf{v}(\cdot\,;\mathbf{s},j)$ defines one of the precisely*

$$\binom{n-1}{j}=\frac{(n-1)!}{j!(n-j-1)!}$$

distinct switching policies of order n, $\mathbf{p}(\pi;j)$, and

(a) *has initial control vector $\mathbf{u}(1,j)=(\boldsymbol{\delta}(j+1),(-1)^{n-j-1}\boldsymbol{\delta}(j))$;*

(b) *switches in the first coordinate when $t=s_1,\cdots,s_j$ and in the second coordinate when $t=s_{j+1},\cdots,s_{n-1}$;*

(c) *terminates with the control $\mathbf{u}(n,j)=((-1)^j\boldsymbol{\delta}(j+1),\boldsymbol{\delta}(j))$.*

(2) *The subset $\Delta(j,T;\pi)^0$ (see (3.14)) of $\Delta(j,T)$ consists of all points generating controls with policy $\mathbf{p}(j,\pi)$ and is, topologically, an open n-cell and its closure $\Delta(j,T;\pi)$ is a closed n-cell.*

(3) *If $\mathbf{p}(j,\pi)=\langle\mathbf{u}(1,j),\mathbf{u}^2,\cdots,\mathbf{u}^{n-1},\mathbf{u}(n,j)\rangle$ and $\mathbf{s}$ is in a boundary cell of $\Delta(j,T;\pi)$, then $\boldsymbol{\lambda}^*(\mathbf{s},j)$ generates control functions whose switching policies are subpolicies of $\mathbf{p}(j,\pi)$, in particular:*

(a) *If $\mathbf{s}$ is in the $(n-1)$-cell $\{\mathbf{s}\in\Delta(j,T;\pi)\colon s_{\pi(1)}=0\}$, then $\boldsymbol{\lambda}^*(\mathbf{s},j)$ generates a control function with switching policy $\mathbf{p}(j,\pi;\mathrm{A})=\langle\mathbf{u}^2,\cdots,\mathbf{u}(n,j)\rangle$ of order $n-1$;*

(b) *If $\mathbf{s}$ is in the $(n-1)$-cell $\{\mathbf{s}\in\Delta(j,T;\pi)\colon s_{\pi(n-1)}=s_n\}$, then $\boldsymbol{\lambda}^*(\mathbf{s},j)$ generates a control function with switching policy $\mathbf{p}(j,\pi;\mathrm{I})=\langle\mathbf{u}(1,j),\cdots,\mathbf{u}^{n-1}\rangle$ of order $n-1$.*

These results will now be used to determine the time-optimal switching policies for small response times of the three example systems introduced in § 1. To simplify

the description we use the notation $\mathbf{u}^1=(1,1)$, $\mathbf{u}^2=(-1,1)$, $\mathbf{u}^3=(-1,-1)$, and $\mathbf{u}^4=(1,-1)$ for the extreme points of $\mathscr{C}^2$. This permits, for example, the description of the policy $\langle \mathbf{u}^i, \mathbf{u}^j, \mathbf{u}^k\rangle$ by, simply, $\langle i,j,k\rangle$.

Example A (continued). For this system, $\delta(0)=\delta(1)=\delta(2)=\delta(3)=1$ and the optimal policies of order 3 are as follows:

$j=0$: $\mathbf{u}(1,0)=\mathbf{u}^1$ and the two optimal policies are $\mathbf{p}$-$\langle 1,4,1\rangle$ and $-\mathbf{p}=\langle 3,2,3\rangle$.

$j=1$: $\mathbf{u}(1,1)=\mathbf{u}^4$ and the optimal policies corresponding to the two permissible permutations are $\mathbf{p}_1=\langle 4,1,2\rangle$, $\mathbf{p}_2=\langle 4,3,2\rangle$ and $-\mathbf{p}_1=\langle 2,3,4\rangle$ and $-\mathbf{p}_2=\langle 2,1,4\rangle$.

$j=2$: $\mathbf{u}(1,2)=\mathbf{u}^1$ and the optimal policies are $\langle 1,2,1\rangle$ and $\langle 3,4,3\rangle$.

Example B (continued). For this system $\delta(0)=\delta(1)=1$ and $\delta(2)=\delta(3)=-1$. This situation leads to the following optimal policies of order 3:

$j=0$: $\mathbf{u}(1,0)=\mathbf{u}^1$ and the optimal policies are $\langle 1,4,1\rangle$ and $\langle 3,2,3\rangle$.

$j=1$: $\mathbf{u}(1,1)=\mathbf{u}^3$ leading to the optimal policies $\langle 3,4,1\rangle$ and $\langle 3,2,1\rangle$ and their negatives $\langle 1,2,3\rangle$ and $\langle 1,4,3\rangle$.

$j=2$: $\mathbf{u}(1,2)=\mathbf{u}^3$ with $\langle 3,4,3\rangle$ and $\langle 1,2,1\rangle$ as optimal policies.

Example C (continued). For the attitude control of the satellite we find $\delta(0)=\delta(1)=\delta(2)=\delta(3)=\delta(4)=1$ and the time-optimal switching policies of order four are as follows:

$j=0$: $\mathbf{u}(1,0)=\mathbf{u}^4$ and the optimal policies are $\langle 2,3,2,3\rangle$ and $\langle 4,1,4,1\rangle$.

$j=1$: $\mathbf{u}(1,1)=\mathbf{u}^1$ with the optimal policies corresponding to the three 1-permissible permutations being $\langle 1,2,3,2\rangle$, $\langle 1,4,3,2\rangle$, and $\langle 1,4,1,2\rangle$ and the negatives $\langle 3,4,1,4\rangle$, $\langle 3,2,1,4\rangle$, and $\langle 3,2,3,4\rangle$.

$j=2$: $\mathbf{u}(1,2)=\mathbf{u}^2$ with $\langle 2,1,2,3\rangle$, $\langle 2,1,4,3\rangle$, and $\langle 2,3,4,3\rangle$ and $\langle 4,3,4,1\rangle$, $\langle 4,3,2,1\rangle$, and $\langle 4,1,2,1\rangle$ as optimal policies.

$j=3$: $\mathbf{u}(2,1)=\mathbf{u}^1$ with the optimal policies $\langle 1,2,1,2\rangle$ and $\langle 3,4,3,4\rangle$.

Remark 3.17. Note that the notation employed to represent the switching policies is such that if two cells have common boundary cells, the associated switching policies have corresponding common subpolicies. For example (Example C, $j=1$), $\langle 1,2,3,2\rangle$ and $\langle 1,4,3,2\rangle$ associated with cells $\{s_1<s_2<s_3<s_4\}$ and $\{s_2<s_1<s_3<s_4\}$ (using an obvious notation), respectively, share the subpolicy $\langle 1,3,2\rangle$ which is the switching policy of control functions associated with the points of the 3-cell $\{s_1=s_2<s_3<s_4\}$. As a result, knowledge of the optimal control functions and their switching policies can be used to analyze the geometry of $\mathbf{K}(T)$ and its switching surfaces.

4. The time-optimal policy complex. In the previous section we identified, for each $j=0,\cdots,n$ and each $\mathbf{s}$ in $\Delta(j,T)$, a time-optimal control function, $\mathbf{v}(\cdot;\mathbf{s},j)$, defined over the interval $0\leqq t\leqq s_n\leqq T$. Because the system (1.1) is normal, $\mathbf{v}(\cdot;\mathbf{s},j)$ is the unique time-optimal control function for a point, $\mathbf{x}(\mathbf{s},j)$, in $\mathbf{K}(T)$. We can identify $\mathbf{x}(\mathbf{s},j)$ by constructing the solution φ to (1.1) with the unknown $\mathbf{x}(\mathbf{s},j)$ as initial data, noting that $\varphi(s_n)=\mathbf{0}$, and solving for $\mathbf{x}(\mathbf{s},j)$ to find

$$\mathbf{x}(\mathbf{s},j)=-\int_0^{s_n} e^{-\mathbf{A}t}\mathbf{B}\mathbf{v}(t;\mathbf{s},j)\,dt. \tag{4.1}$$

If the time-optimal switching policy of $\mathbf{v}(\cdot;\mathbf{s},j)$ is $\mathbf{p}(j,\pi)=\langle \mathbf{u}^1,\mathbf{u}^2,\cdots,\mathbf{u}^n\rangle$ (or a subpolicy of $\mathbf{p}(j,\pi)$), then

$$\begin{aligned}\mathbf{x}(\mathbf{s},j)&=-\int_0^{s_{\pi(1)}} e^{-\mathbf{A}t}\mathbf{B}\,dt\cdot\mathbf{u}^1-\int_{s_{\pi(1)}}^{s_{\pi(2)}} e^{-\mathbf{A}t}\mathbf{B}\,dt\cdot\mathbf{u}^2-\cdots-\int_{s_{\pi(n-1)}}^{s_n} e^{-\mathbf{A}t}\mathbf{B}\,dt\cdot\mathbf{u}^n\\&=-\sum_{k=1}^{n-1}\int_0^{s_{\pi(k)}} e^{-\mathbf{A}t}\mathbf{B}\,dt\cdot(\mathbf{u}^k-\mathbf{u}^{k+1})-\int_0^{s_n} e^{-\mathbf{A}t}\mathbf{B}\,dt\cdot\mathbf{u}^n.\end{aligned} \tag{4.2}$$

The relationship is simplified by defining

$$\mathbf{g}^i(\tau) \equiv -\int_0^\tau e^{-\mathbf{A}t}\mathbf{b}^i\,dt, \qquad i=1,2 \tag{4.3}$$

and noting that for $1 \leqq \pi(k) \leqq j$, the switch from $\mathbf{u}^k$ to $\mathbf{u}^{k+1}$ at time $t = s_{\pi(k)}$ involves the first coordinate and, from Lemma 3.12(1),

$$\mathbf{u}^k - \mathbf{u}^{k+1} = \begin{bmatrix} \pm 2\boldsymbol{\delta}(j+1) \\ 0 \end{bmatrix}$$

("+" if $\pi(k)=1$, "−" if $\pi(k)=2, \cdots$), while if $j+1 \leqq \pi(k) \leqq n-1$, the switch involves the second coordinate and

$$\mathbf{u}^k - \mathbf{u}^{k+1} = \begin{bmatrix} 0 \\ \pm(-1)^{n-j-1}2\boldsymbol{\delta}(j) \end{bmatrix}$$

("+" if $\pi(k)=j+1$, "−" if $\pi(k)=j+2, \cdots$). Then, in terms of

$$\mathbf{G}^1(\mathbf{s}; j) = \mathbf{g}^1(s_1) - \mathbf{g}^1(s_2) + \cdots + (-1)^{j-1}\mathbf{g}^1(s_j) + \frac{(-1)^j\mathbf{g}^1(s_n)}{2} \tag{4.4}$$

and

$$\mathbf{G}^2(\mathbf{s}; j) = \mathbf{g}^2(s_{j+1}) - \mathbf{g}^2(s_{j+2}) + \cdots + (-1)^{n-j-2}\mathbf{g}^2(s_{n-1}) + \frac{(-1)^{n-j-1}\mathbf{g}^2(s_n)}{2}$$

we have

$$\mathbf{x}(\mathbf{s}, j) = 2\boldsymbol{\delta}(j+1)\mathbf{G}^1(\mathbf{s}; j) + (-1)^{n-j-1}2\boldsymbol{\delta}(j)\mathbf{G}^2(\mathbf{s}; j), \tag{4.5}$$

which is independent of π, valid for all $\mathbf{s}$ in $\boldsymbol{\Delta}(j, T)$ and defines a C^∞ mapping $\mathbf{s} \mapsto \mathbf{x}(\mathbf{s}, j)$ of $\boldsymbol{\Delta}(j, T)$ onto a subset, $\mathbf{D}(j, T)$, of $\mathbf{K}(T)$. $\mathbf{D}(j, T)$, being the continuous image of a compact set, is compact and consists of all points $\mathbf{x}$ in $\mathbf{K}(T)$ which are time-optimally controlled to $\mathbf{0}$ by control functions $\mathbf{v}(\cdot; \mathbf{s}, j)$ with $\mathbf{s} \in \boldsymbol{\Delta}(j, T)$. Because of its cell structure, which is developed below, $\mathbf{D}(j, T)$ is called a **time-optimal policy complex.**

The map (4.5) is the composition $\mathbf{s} \mapsto \mathbf{v}(\cdot; \mathbf{s}, j) \mapsto \mathbf{x}(\mathbf{s}, j)$ and, while the latter is injective because of the normality of the system, the former is necessarily injective only on the interior of $\boldsymbol{\Delta}(j, T)$. In fact, it will not be injective on the boundary of $\boldsymbol{\Delta}(j, T)$ if either j or $n-j-1$ is greater than 1. Because of the central role these issues play in the sections to follow, it will be beneficial to explore an example in detail. Toward that end we turn to the case "$j=1$" for the fourth-order system, Example C, which provides a representative example.

For Example C and $j=1$, $\boldsymbol{\Delta}(1, T) = \{\mathbf{s} \in \mathbb{R}^4 : 0 \leqq s_1 \leqq s_4 \leqq T,\ 0 \leqq s_2 \leqq s_3 \leqq s_4 \leqq T\}$ and (recall that all $\boldsymbol{\delta}(i) = 1$ for all i)

$$\mathbf{x}(\mathbf{s}, 1) = 2[\mathbf{g}^1(s_1) - \mathbf{g}^1(s_4)/2] + 2[\mathbf{g}^2(s_2) - \mathbf{g}^2(s_3) + \mathbf{g}^2(s_4)/2] \tag{4.6}$$

maps $\boldsymbol{\Delta}(1, T) \mapsto \mathbf{D}(1, T) \subset \mathbf{K}(T) \subset \mathbb{R}^4$. The interior of $\boldsymbol{\Delta}(1, T)$ is composed of the three open 4-cells (using obvious notation) $\boldsymbol{\Delta}(1, T; \pi_0)^0 = \{0 < s_1 < s_2 < s_3 < s_4 < T\}$ (π_0 is the identity permutation), $\boldsymbol{\Delta}(1, T; \pi_1)^0 = \{0 < s_2 < s_1 < s_3 < s_4 < T\}$, and $\boldsymbol{\Delta}(1, t; \pi_2)^0 = \{0 < s_2 < s_3 < s_1 < s_4 < T\}$ and their shared open 3-cell boundaries $\{0 < s_1 = s_2 < s_3 < s_4 < T\}$ and $\{0 < s_2 < s_1 = s_3 < s_4 < T\}$.

Remark 4.7. In this context "open 3-cell (open 2-cell, etc.)" indicates a cellular homeomorphic image of the standard open 3-cell (open 2-cell, etc.) not necessarily, an open set in the topology of the space (see [16]).

Through the outer normal vectors $\boldsymbol{\lambda}(\cdot, 1)$ (see (3.3)), the points of these cells are associated with time-optimal control functions having the 4-policies (recall the notation from the previous section) $\langle 1, 2, 3, 2\rangle$, $\langle 1, 4, 3, 2\rangle$, and $\langle 1, 4, 1, 2\rangle$ and the 3-policies $\langle 1, 3, 2\rangle$ and $\langle 1, 4, 2\rangle$, respectively.

The boundary of $\boldsymbol{\Delta}(1, T)$ contains the set $\{\mathbf{s} \in \boldsymbol{\Delta}(1, T)\colon s_4 = T\}$, which $\mathbf{x}(\cdot, 1)$ maps into the T-isochrone, and the following cell complexes each composed of one or two open 3-cells:

Cell complex	Optimal policy
$\{0 = s_1, 0 < s_2 < s_3 < s_4 < T\}$	$\langle 2, 3, 2\rangle$
$\{s_1 = s_4 < T, 0 < s_2 < s_3 < s_4 < T\}$	$\langle 1, 4, 1\rangle$
$\{0 < s_1 < s_4 < T, 0 = s_2 < s_3 < s_4 < T\}$	$\langle 4, 3, 2\rangle$ and $\langle 4, 1, 2\rangle$
$\{0 < s_1 < s_4 < T, 0 < s_2 < s_3 = s_4 < T\}$	$\langle 1, 2, 3\rangle$ and $\langle 1, 4, 3\rangle$
$\{0 < s_1 < s_4 < T, 0 < s_2 = s_3 < s_4 < T\}$	$\langle 1, 2\rangle$

The last three complexes of this list each contain two open 3-cells (e.g., corresponding to $s_1 < s_3$ and $s_3 < s_1$ in the first instance) which, in the first two cases, determine two optimal 3-policies. In the last instance the points of the two open 3-cells determine a single 2-policy which is also the policy associated with the open 2-cell $\{0 < s_1 < s_2 = s_3 = s_4 < T\}$. Other boundary cells of $\boldsymbol{\Delta}(1, T)$ determine additional 2- and 1-policies. For example:

Cell complex	Optimal policy
$\{0 = s_1 < s_2 < s_3 = s_4 < T\}$	$\langle 2, 3\rangle$
$\{0 = s_1 = s_2 < s_3 < s_4 < T\}$	$\langle 3, 2\rangle$
$\{0 = s_2 < s_1 < s_3 = s_4 < T\}$	$\langle 4, 3\rangle$
$\{0 = s_2 < s_1 = s_3 = s_4 < T\}$	$\langle 4\rangle$

The connection between shared boundary cells of lower dimension and shared subpolicies (see Remark 3.17) is clearly evident from these examples. Furthermore, the "shared subpolicies" become, via (4.5), shared boundary cells of the policy complexes. This illustrates the geometric insight provided by the concept of "switching policy."

Now, as noted above, there is a one-to-one correspondence between control functions and points of $\mathbf{K}(T)$ because of the assumed normality of the system. However, as the examples above clearly demonstrate, the correspondence $\mathbf{s} \mapsto \mathbf{x}(\mathbf{s}, 1)$ is, in general, many to one when $\mathbf{s}$ lies on the boundary of $\boldsymbol{\Delta}(1, T)$. Before specifically addressing the many-to-one feature, let us consider the mapping in more detail. From (4.4) and (4.5), the Jacobian matrix of (4.6) (in column form) is

$$\frac{\partial \mathbf{x}(\mathbf{s}, 1)}{\partial \mathbf{s}} = [-2\mathbf{E}^1(s_1), -2\mathbf{E}^2(s_2), 2\mathbf{E}^2(s_3), \mathbf{E}^1(s_4) - \mathbf{E}^2(s_4)] \tag{4.8}$$

with the determinant

$$\begin{aligned} \det\left[\frac{\partial \mathbf{x}(\mathbf{s}, 1)}{\partial \mathbf{s}}\right] &= 8[\boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^1(s_4) - \boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^2(s_4)] \\ &= 8[|\boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^1(s_4)| + |\boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^2(s_4)|], \end{aligned} \tag{4.9}$$

given the signs assumed by the determinants $\boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^i(t)$ for $t > \max(s_1, s_3)$ (shown in (3.9b)). This result confirms the local injectivity of the mapping on the interior of $\boldsymbol{\Delta}(1, T)$ as long as both determinants, $\boldsymbol{\lambda}(\mathbf{s}, 1) \wedge \mathbf{E}^i(t)$, $i = 1, 2$, do not vanish identically. More important, it leads to a proof that on the relative interior of a boundary k-cell,

that is, an open k-cell (in this case $k = 1, 2$, or 3) corresponding to an optimal k-policy, the Jacobian matrix has rank k, and therefore, the image of this open k-cell in $\mathbf{D}(1, T)$ is also an open k-cell (note Remark (4.7)).

This example is typical of the general case. Calculations similar to those of (4.8) and (4.9) lead to the same conclusions concerning the Jacobian of (4.5) and show that this C^∞ mapping is injective on the interior of $\mathbf{\Delta}(j, T)$ and on any open k-cell associated with a time-optimal k-policy. However, the mapping is not injective on some cells defined by equalities of the form $s_i = s_m$, where $1 \leqq i,\ m \leqq j$, or $j+1 \leqq i,\ m \leqq n-1$. Examination of (4.5) shows that such a cell of dimension, say, m is associated with the same optimal k-policy, with $k < m$, as some open k-cell. Thus the two cells are, insofar as time-optimal controls are concerned, equivalent, since their points $\mathbf{s}$ map into the same open k-cell in $\mathbf{D}(j, T)$. The following proposition, proven in [23], uses the concepts and results of [16] to formalize these relationships.

PROPOSITION 4.10. *The equivalence relation on* $\mathbf{\Delta}(j, T)$ *defined by*

$$\mathbf{R}(j) \equiv \{(\mathbf{s}^1, \mathbf{s}^2)\colon \mathbf{s}^1, \mathbf{s}^2 \in \mathbf{\Delta}(j, T), \mathbf{v}(\cdot; \mathbf{s}^1, j) = \mathbf{v}(\cdot; \mathbf{s}^2, j)\}$$

is a cellular equivalence relation. Also,

(1) *With the usual topology on the cell complex* $\mathbf{\Delta}(j, t)$, *the quotient space* $\mathbf{\Delta}(j, T)/\mathbf{R}(j)$ *is a normal* CW-*complex*;

(2) *The map* $\mathbf{x}^*\colon \mathbf{\Delta}(j, T)/\mathbf{R}(j) \mapsto \mathbf{D}(j, T) \subset \mathbf{K}(T)$ *based on* (4.5) *is a cellular homeomorphism, and therefore, the* **time-optimal policy complex** $\mathbf{D}(j, T)$ *is also a normal* CW-*complex.*

This proposition provides a clear picture of the policy complex $\mathbf{D}(j, T)$. Its interior is, essentially, that of $\mathbf{\Delta}(j, T)$ while its boundary contains homeomorphic images of some of the $(n-1)$-dimensional boundary cells of $\mathbf{\Delta}(j, T)$. Other boundary cells of the latter cell complex are mapped by $\mathbf{x}(\cdot, j)$ into a single cell of lower dimension. This situation is especially clear in the third-order Examples A and B for $j = 0$ or 3. It is obvious that the face of the tetrahedron $\mathbf{\Delta}(0, T)$ defined by $s_1 = s_2$ is mapped by $\mathbf{x}(\mathbf{s}, 0) = \boldsymbol{\delta}(1)\mathbf{g}^1(s_3) + 2\boldsymbol{\delta}(0)[\mathbf{g}^2(s_1) - \mathbf{g}^2(s_2) + \mathbf{g}^2(s_3)/2]$ into a curve which is the homeomorphic image of the cell $\{\mathbf{s} \in \mathbf{\Delta}(0, T)\colon 0 = s_1 = s_2 \leqq s_3 \leqq T\}$.

This qualitative view is, however, not sufficient; we need considerably more detail in order to reconstruct the time-optimal flow on $\mathbf{K}(T)$. Much of the necessary information is provided by identification of the cells of $\mathbf{D}(j, T)$ having dimension n and $n-1$.

The only open n-cells of the complex $\mathbf{D}(j, T)$ are the images of the $\binom{n-1}{j}$ cells $\mathbf{\Delta}(j, T; \pi)^0$, defined by the j-permissible permutations π. Images of open $(n-1)$-cells of the form $\{\mathbf{s} \in \mathbf{\Delta}(j, T)^0\colon s_i = s_k, 1 \leqq i \leqq j < k \leqq n-1, s_\alpha \neq s_\beta, \alpha, \beta \neq i, j\}$ lie in the interior of $\mathbf{D}(j, T)$ and form the joint boundaries of the open n-cells. These cells play no particular role in the following. That is far from the case for the open $(n-1)$-cells forming the boundary of $\mathbf{D}(j, T)$ relative to $\mathbf{K}(T)$ (that is, ignoring the portion of the boundary contained in the T-isochrone). These cells occur in four groups and are the images under (4.5) of the cell complexes of $\mathbf{\Delta}(j, T)$ defined by $s_1 = 0$, by $s_j = s_n$, by $s_{j+1} = 0$, and by $s_{n-1} = s_n$. These complexes will be denoted by $\mathbf{D}(j|s_1 = 0)$, $\mathbf{D}(j|s_j = s_n)$, etc. The first two of these complexes correspond to those permutations in which the first, or last, switch occurs in the first coordinate; the latter two complexes corresponding to those permutations in which the first or last switch occurs in the second coordinate.

$\mathbf{D}(j|s_1 = 0)$ and $\mathbf{D}(j|s_j = s_n)$ each contain $\binom{n-2}{j-1}$ open $(n-1)$-cells corresponding to each j-permissible permutation with $\pi(1) = 1$ (so that $s_1 < s_{j+1}$) for the first complex and to the permutations with $\pi(n-1) = j$ (so that $s_{n-1} < s_j$) for the second. $\mathbf{D}(j|s_{j+1} = 0)$

and $\mathbf{D}(j|s_{n-1}=s_n)$ each contain $\binom{n-2}{j}$ open $(n-1)$-cells corresponding to those π satisfying $\pi(1)=j+1$ in the first instance, or $\pi(n-1)=n-1$ in the second.

These relationships can be made clearer by defining

$$\mathbf{D}(j, T; \pi)=\operatorname{clos}[\mathbf{x}(\mathbf{\Delta}(j, T; \pi), j)], \tag{4.11}$$

which is a closed n-cell and consists of all points of $\mathbf{K}(T)$ whose time-optimal control functions have switching policies that are subpolicies of $\mathbf{p}(j, \pi)$ (see 3.16). In particular, the open $(n-1)$-cell corresponding to the subpolicy defined as $\mathbf{p}(j, \pi, \mathrm{A})$ is a cell of the complex $\mathbf{D}(j|s_1=0)$ if $\pi(1)=1$, and of the complex $\mathbf{D}(j|s_{j+1}=0)$ if $\pi(1)=j+1$. Similarly, the open $(n-1)$-cell associated with the subpolicy $\mathbf{p}(j, \pi, \mathrm{I})$ is in $\mathbf{D}(j|s_j=s_n)$ or $\mathbf{D}(j|s_{n-1}=s_n)$ if $\pi(n-1)=j$ or $n-1$, respectively.

5. The time-optimal flow as a local semidynamical system. In this section we examine the time-optimal trajectories on the policy cell complexes $\pm\mathbf{D}(j, T)$, $j=0, \cdots, n$, of the previous sections from the perspective provided by the theory of semidynamical systems (see [1]). In the process we show that the policies $\mathbf{p}(j, \pi, \mathrm{A})$ and $\mathbf{p}(j, \pi, \mathrm{I})$ of order $n-1$ correspond, indeed, to attracting and invariant cells of the local semidynamical system defined by the time-optimal trajectories on $\mathbf{D}(j, T)\backslash\{\mathbf{0}\}$.

A point $\mathbf{x}(\mathbf{s}, j)$ in the interior of $\mathbf{D}(j, T)$ lies either in the interior of a unique closed n-cell $\mathbf{D}(j, T, \pi)$ or, if its optimal control function has a policy of order $<n$, in the boundary of two or more such n-cells. For the moment let us assume the former (i.e., that the coordinates of $\mathbf{s}$ are distinct) so that the time-optimal control for this point has the n-policy $\mathbf{p}(j, \pi)\equiv\langle\mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^{n-1}, \mathbf{u}^n\rangle$ with switches occurring in the sequence $t=s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n-1)}$. Define

$$\mathbf{y}(s_{\pi(1)}, s_{\pi(2)}, \cdots, s_{\pi(n-1)}, s_n)\equiv\mathbf{x}(s_1, s_2, \cdots, s_{n-1}, s_n, j). \tag{5.1}$$

Then, since $\mathbf{x}(\mathbf{s}, j)$ was found by "backing out from $\mathbf{0}$" in (4.1), the time-optimal trajectory from $\mathbf{x}(\mathbf{s}, j)$ to $\mathbf{0}$ is easily seen (using (4.2)) to be described by

$$\begin{aligned}
\varphi(t) &= \mathbf{y}(s_{\pi(1)}-t, s_{\pi(2)}-t, \cdots, s_{\pi(n-1)}-t, s_n-t), && 0\leqq t<s_{\pi(1)},\\
&= \mathbf{y}(0, s_{\pi(2)}-t, \cdots, s_{\pi(n-1)}-t, s_n-t), && s_{\pi(1)}\leqq t<s_{\pi(2)},\\
&= \mathbf{y}(0, 0, s_{\pi(3)}-t, \cdots, s_{\pi(n-1)}-t, s_n-t), && s_{\pi(2)}\leqq t<s_{\pi(3)},\\
&\ \ \vdots && \ \ \vdots\\
&= \mathbf{y}(0, 0, \cdots, 0, s_{\pi(k)}-t, \cdots, s_{\pi(n-1)}-t, s_n-t), && s_{\pi(k-1)}\leqq t<s_{\pi(k)},\\
&\ \ \vdots && \ \ \vdots\\
&= \mathbf{y}(0, 0, \cdots, 0, s_n-t), && s_{\pi(n-1)}\leqq t<s_n,\\
&= \mathbf{0}, && s_n\leqq t\leqq T.
\end{aligned} \tag{5.2}$$

This representation of the time-optimal trajectories can be extended to all of $\mathbf{D}(j, T)$ and is easily seen to define a semidynamical system [1] on this complex. It is clear that each point of the T-isochrone ($s_n=T$) is a "start-point" of the flow (i.e., is not "downstream" from another point of the complex). Furthermore, as the following proposition shows, the semidynamical system is strongly related to the cell structure of the policy complex.

PROPOSITION 5.3. *Let $\mathbf{x}^0=\mathbf{x}(\mathbf{s}, j)$ be a point of $\mathbf{D}(j, T)$ and suppose that the switching policy of the time-optimal control function for $\mathbf{x}^0$, $\mathbf{v}(\cdot; \mathbf{s}, j)$, is $\mathbf{p}(j, \pi)$ or a subpolicy of $\mathbf{p}(j, \pi)$. Let $\varphi(t)$, $0\leqq t<s_n$ describe the time-optimal trajectory from $\mathbf{x}^0$ to $\mathbf{0}$ (see (5.2)) and $\mathbf{D}(j, \pi, \alpha)$ be the image under $\mathbf{x}(\cdot, j)$ of the $(n-\alpha)$-dimensional cell of $\mathbf{\Delta}(j, T; \pi)$ defined by $s_{\pi(1)}=s_{\pi(2)}=\cdots=s_{\pi(\alpha)}=0$. Then*

(1) *$\mathbf{D}(j, \pi, \alpha)$ is a closed cell of $\mathbf{D}(j, T)$ of dimension $n-\alpha$;*

(2) *If* $s_{\pi(1)} > 0$, $\varphi(t)$ *lies in the interior of* $\mathbf{D}(j, T)$ *for* $0 \leqq t < s_{\pi(1)}$;

(3) *If* $s_{\pi(\alpha)} < s_{\pi(\alpha+1)}$, $\varphi(t)$ *lies in the relative interior of* $\mathbf{D}(j, \pi, \alpha)$ *for* $s_{\pi(\alpha)} < t < s_{\pi(\alpha+1)}$;

(4) *If* $\mathbf{s}$ *has distinct coordinates, then* $\varphi(t)$ *successively traverses the relative interior of each cell of the sequence* $\mathbf{D}(j, T), \mathbf{D}(j, \pi, 1), \mathbf{D}(j, \pi, 2), \cdots, \mathbf{D}(j, \pi, n-1)$ *as* t *traverses the interval* $[0, s_n]$.

Proof. The first result follows from continuity of the mapping $\mathbf{x}(\cdot, j)$ and calculations based on its Jacobian analogous to (4.8) and (4.9). The other assertions follow directly from (5.2). □

A subset M is an attracting set of a semidynamical system on a set X if the set of points of X that are attracted to M form a neighborhood of M (the notion of "attraction" is clear for the time-optimal flow considered here; see [1] for precise technical considerations). Thus, in considering the time-optimal flow on $\mathbf{D}(j, T)$, it is clear that any set containing $\mathbf{0}$ will be an attractor of the flow. This is, obviously, of little use in studying the time-optimal system. The notion of attraction, however, becomes much more discriminating when we consider the local semidynamical system defined on $\mathbf{D}(j, T)\backslash\{\mathbf{0}\}$ by the time-optimal flow. In this instance (5.2) and (5.3) show that the cells $\mathbf{D}(j, \pi, 1)$ and their subcells are attractors of the flow restricted to $\mathbf{D}(j, T; \pi)$. When all j-permissible permutations are considered, these cells form the **attracting cell complexes** denoted by $\pm\mathbf{D}(j|s_1 = 0)$ and $\pm\mathbf{D}(j|s_{j+1} = 0)$ in § 4.

A subset N is a positively invariant set of a semidynamical system on a set X if the trajectories starting in N remain in N. It is called negatively invariant if its complement is positively invariant, and invariant if it is both positively and negatively invariant. In applying these notions to the time-optimal flow on $\mathbf{D}(j, T)$, we note that the cellular nature of the flow makes each closed cell of the nested sequences $\mathbf{D}(j, \pi, 1)$, $\mathbf{D}(j, \pi, 2), \cdots, \mathbf{D}(j, \pi, n-1)$ positively invariant, but not invariant, cells. However, as (5.2) shows, the cells defined by $s_{\pi(n-1)} = s_n$ are both positively and negatively invariant. These $(n-1)$-dimensional cells form the **invariant cell complexes** previously denoted by $\pm\mathbf{D}(j|s_j = s_n)$ and $\pm\mathbf{D}(j|s_{n-1} = s_n)$. These results are summarized in Proposition 5.4.

PROPOSITION 5.4. *The local semidynamical system defined on* $\mathbf{D}(j, T)\backslash\{0\}$ *by the time-optimal flow characterizes the boundary cells of this policy complex as follows*:

(1) *The portion of the T-isochrone contained in the policy complex consists of "start points" of the time-optimal flow*;

(2) *The* $(n-1)$*-dimensional cell complexes* $\mathbf{D}(j|s_1 = 0)$ *and* $\mathbf{D}(j|s_{j+1} = 0)$ *are* **attracting sets** *of the time-optimal flow*;

(3) *The* $(n-1)$*-dimensional cell complexes* $\mathbf{D}(j|s_j = s_n)$ *and* $\mathbf{D}(j|s_{n-1} = s_n)$ *are* **invariant sets** *of the time-optimal flow.*

6. The assembly of K(*T*). In the previous section we identified the $n+1$ policy complexes $\mathbf{D}(j, T)$, $j = 0, 1, \cdots, n$ and the complexes composed of their $(n-1)$-dimensional boundary cells $\mathbf{D}(j|s_1 = 0), \cdots, \mathbf{D}(j|s_{n-1} = s_n)$, $j = 0, 1, \cdots, n$. Due to the symmetry of the system—$-\mathbf{v}(\cdot; \mathbf{s}, j)$ is a time-optimal control function if $\mathbf{v}(\cdot; \mathbf{s}, j)$ is a time-optimal control function—we also have their negatives $-\mathbf{D}(j, T)$, etc. In this section we show how this "n-dimensional jigsaw puzzle" assembles to form $\mathbf{K}(T)$. The basic step in the assembly process involves showing that each boundary $(n-1)$-cell of, for example, $\mathbf{D}(j, T)$ is also a boundary cell for precisely one of the four complexes $\pm\mathbf{D}(j-1, T)$ or $\pm\mathbf{D}(j+1, T)$. This we accomplish by direct calculation. The remainder of the construction relies on a technical lemma that shows the union of all the policy complexes to be $\mathbf{K}(T)$. In this process we are providing another proof of Yeung's result on the minimal controllability of strictly normal systems of the type (1.1) [29]

and providing a great deal more information on the time-optimal flow which will be of use in later sections.

LEMMA 6.1 (Structure lemma). *Let* $r=2$ *and system* (1.1) *be strictly normal and define*

$$\boldsymbol{\gamma}(k)=\boldsymbol{\delta}(k-1)\cdot\boldsymbol{\delta}(k+1),\qquad k=1,\cdots,n-1. \tag{6.2}$$

Then, for $1\leqq j\leqq n-1$:

(1) *If* $\boldsymbol{\gamma}(j)=1$

1) $\mathbf{D}(j|s_1=0)=-\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_{n-1}=s_n)$,

2) $\mathbf{D}(j|s_j=s_n)=\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_j=0)$;

(2) *If* $\boldsymbol{\gamma}(j)=-1$

1) $\mathbf{D}(j|s_1=0)=\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_j=0)$,

2) $\mathbf{D}(j|s_j=s_n)=-\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_{n-1}=s_n)$;

For $0\leqq j\leqq n-2$:

(3) *If* $\boldsymbol{\gamma}(j+1)=1$

1) $\mathbf{D}(j|s_{j+1}=0)=\boldsymbol{\delta}(j+1)\cdot\boldsymbol{\delta}(j)\cdot\mathbf{D}(j+1|s_{j+1}=s_n)$,

2) $\mathbf{D}(j|s_{n-1}=s_n)=-\boldsymbol{\delta}(j+1)\cdot\boldsymbol{\delta}(j)\cdot\mathbf{D}(j+1|s_1=0)$;

(4) *If* $\boldsymbol{\gamma}(j+1)=-1$

1) $\mathbf{D}(j|s_{j+1}=0)=\boldsymbol{\delta}(j+1)\cdot\boldsymbol{\delta}(j)\cdot\mathbf{D}(j+1|s_1=0)$,

2) $\mathbf{D}(j|s_{n-1}=s_n)=-\boldsymbol{\delta}(j+1)\cdot\boldsymbol{\delta}(j)\cdot\mathbf{D}(j+1|s_{j+1}=s_n)$.

Proof. As noted above, the proofs of these relationships are computational. Because, from (4.4), $\mathbf{G}^1(\mathbf{s};j)$ depends only on $s_1,\cdots,s_j,s_n$ and $\mathbf{G}^2(\mathbf{s};j)$ depends only on $s_{j+1},\cdots,s_{n-1},s_n$, we have the relationships

$$\mathbf{G}^1(0,s_2,\cdots,s_j,s_n;j)=-\mathbf{G}^1(s_2,\cdots,s_j,s_n;j-1),$$

$$\mathbf{G}^1(s_1,\cdots,s_{j-1},s_n,s_n;j)=\mathbf{G}^1(s_1,\cdots,s_{j-1},s_n;j-1)$$

and

$$\mathbf{G}^2(0,s_{j+2},\cdots,s_{n-1},s_n;j)=-\mathbf{G}^2(s_{j+2},\cdots,s_{n-1},s_n;j+1),$$

$$\mathbf{G}^2(s_{j+1},\cdots,s_{n-2},s_n,s_n;j)=\mathbf{G}^2(s_{j+1},\cdots,s_{n-2},s_n;j+1).$$

Now, using the first two of these equations and (4.5), we find that

$$\begin{aligned}\boldsymbol{\delta}(j-1)\cdot\mathbf{x}(\mathbf{s};j)|_{s_1=0}&=-2\boldsymbol{\gamma}(j)\mathbf{G}^1(s_2,\cdots,s_j,s_n;j-1)\\&\quad+(-1)^{n-j-1}2\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\mathbf{G}^2(s_{j+1},\cdots,s_n;j),\end{aligned} \tag{6.3a}$$

$$\begin{aligned}\boldsymbol{\delta}(j-1)\cdot\mathbf{x}(\mathbf{s};j)|_{s_j=s_n}&=2\boldsymbol{\gamma}(j)\mathbf{G}^1(s_1,\cdots,s_{j-1},s_n;j-1)\\&\quad+(-1)^{n-j-1}2\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\mathbf{G}^2(s_{j+1},\cdots,s_n;j),\end{aligned} \tag{6.3b}$$

$$\begin{aligned}\boldsymbol{\delta}(j)\cdot\mathbf{x}(\mathbf{s};j-1)|_{s_j=0}&=2\mathbf{G}^1(s_1,\cdots,s_{j-1},s_n;j-1)\\&\quad+(-1)^{n-j-1}2\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\mathbf{G}^2(s_{j+1},\cdots,s_n;j),\end{aligned} \tag{6.4a}$$

$$\begin{aligned}\boldsymbol{\delta}(j)\cdot\mathbf{x}(\mathbf{s};j-1)|_{s_{n-1}=s_n}&=2\mathbf{G}^1(s_1,\cdots,s_{j-1},s_n;j-1)\\&\quad+(-1)^{n-j}2\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\mathbf{G}^2(s_j,\cdots,s_{n-2},s_n;j).\end{aligned} \tag{6.4b}$$

Now, if $\boldsymbol{\gamma}(j)=1$, we see that the image of the $(n-1)$-cells making up the set $(\mathbf{s}\in\boldsymbol{\Delta}(j,T)\colon s_1=0\}$ under the map (6.3a) is exactly the same as the image of the $(n-1)$-cells of the set $\{\mathbf{s}\in\boldsymbol{\Delta}(j-1,T)\colon s_{n-1}=s_n\}$ under the negative of the map (6.4b).

Similarly, the image of $\{\mathbf{s}\in\boldsymbol{\Delta}(j, T)\colon s_j = s_n\}$ under the map of (6.3b) is equal to the image of $\{\mathbf{s}\in\Delta(j-1, T)\colon s_j = 0\}$ under the map of (6.4a). Thus, when $\boldsymbol{\gamma}(j)=1$, $\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j|s_1=0) = -\boldsymbol{\delta}(j)\cdot\mathbf{D}(j-1|s_{n-1}=s_n)$, and $\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j|s_j=s_n)=\boldsymbol{\delta}(j)\cdot\mathbf{D}(j-1|s_j=0\}$. This proves (1(1)) and (1(2)) of the lemma. The assertions (2(1)) and (2(2)) also follow from (6.3) and (6.4) under the assumption that $\boldsymbol{\gamma}(j)=-1$.

In the same fashion we can prove the remaining assertions of the lemma by considering the mappings $\boldsymbol{\delta}(j+1)\cdot\mathbf{x}(\cdot, j)$ and $\boldsymbol{\delta}(j)\cdot\mathbf{x}(\cdot, j+1)$ restricted to the appropriate cells of $\boldsymbol{\Delta}(j, T)$ and $\boldsymbol{\Delta}(j+1, T)$. □

Now let

$$\mathbf{K}_1 = \bigcup_{j=0}^{n-1}[\mathbf{D}(j, T)\cup(-\mathbf{D}(j, T))].$$

$\mathbf{K}_1$ is a closed cell complex contained in $\mathbf{K}(T)$ and, from (6.1), since $\mathbf{D}(0, T)$ shares an $(n-1)$-cell with $\mathbf{D}(1, T)$ and $-\mathbf{D}(1, T), \cdots, \mathbf{D}(n-1, T)$ shares an $(n-1)$-cell with $\mathbf{D}(n, T)$ and $-\mathbf{D}(n, T)$, it is clear that $\mathbf{K}_1$ is a connected complex which is symmetric with respect to $\mathbf{0}$. Furthermore, its interior $\mathbf{K}_1^0$ is also connected. To see this, let $\mathbf{x}^0$ be a point of the relative interior of one of the boundary $(n-1)$-cells of, say, $\mathbf{D}(j, T)$. Then the map (4.5) is locally a smooth homeomorphism and so, near $\mathbf{x}^0$, $\mathbf{D}(j, T)$ is a smooth n-dimensional manifold with boundary which contains a half-neighborhood of $\mathbf{x}^0$ intersecting the boundary $(n-1)$-cell in a relatively open set. The same result is true for the other complex, say, $\mathbf{D}(j-1, T)$, which shares the cell containing $\mathbf{x}^0$ with $\mathbf{D}(j, T)$. The union of the two half-neighborhoods forms a neighborhood of $\mathbf{x}^0$ contained in $\mathbf{D}(j, T)\cup\mathbf{D}(j-1, T)$ which, clearly, has a connected interior. This, given the relationships of Lemma 6.1, proves that $\mathbf{K}_1$ is a closed symmetric cell complex with nonempty connected interior.

THEOREM 6.5 (Cellular Decomposition Theorem). *Let $r=2$ and system (1.1) be strictly normal and let T be sufficiently small and positive. Then*

$$\mathbf{K}(T) = \bigcup_{j=0}^{n-1}[\mathbf{D}(j, T)\cup(-\mathbf{D}(j, T))].$$

Proof. $\mathbf{K}(T)$ is a connected convex set with nonempty interior, $\mathbf{K}(T)^0$. Thus $\mathbf{K}(T)^0\backslash\mathbf{K}_1$ is either empty or contains a point $\mathbf{x}^1$. In the first case the result is proved. In the second instance, let $\mathbf{x}^2$ be a point of the interior of $\mathbf{K}_1$. Since both $\mathbf{x}^1$ and $\mathbf{x}^2$ lie in the interior of the connected set $\mathbf{K}(T)^0$, there exists a curve lying in $\mathbf{K}(T)^0$ that connects them. The cells of $\mathbf{K}_1$ of dimension $n-2$ or less (the cells of the $(n-2)$-*skeleton* of $\mathbf{K}_1$; see [16]) are a relatively negligible part of the complex (and of $\mathbf{K}(T)$) and it is reasonable to suspect that the curve connecting these two points can be chosen to avoid all such cells. This is, in fact, possible by a general position argument [8], [28]. Thus, if $\mathbf{z}$ is the "last" point of the connecting curve lying in the closed set $\mathbf{K}_1$, it must lie in the relative interior of some cell of dimension $n-1$. This, however, is impossible, since we have just shown that the points of the relative interiors of all such $(n-1)$-cells are in the interior of $\mathbf{K}_1$. This contradiction proves the proposition. □

This result shows that the control functions $\pm\mathbf{v}(\cdot; j, \mathbf{s})$, $\mathbf{s}\in\boldsymbol{\Delta}(j, T)$, $j=0, \cdots, n-1$, we have identified are sufficient to control all points of $\mathbf{K}(T)$, for sufficiently small T, to $\mathbf{0}$ time-optimally. In the sections to follow we shall use this result to explore the time-optimal flow, the switching surface structure and the character of the time-optimal feedback function.

7. Time-optimal feedback control and regular synthesis. In the previous section we proved that $\mathbf{K}(T)$ is the union of the policy cell complexes $\pm\mathbf{D}(j, T)$, $j=0, \cdots, n-1$. By the normality of the system, time-optimal controls are unique. It is clear, therefore,

that the unique control function for a point lying in the intersection of two, or more, policy complexes has a policy that is a subpolicy of one or more of the n-policies associated with the original complexes. Thus, if we define the time-optimal feedback control function on each of the component n-complexes, $\pm\mathbf{D}(j, T)$, $j=0, \cdots, n-1$, it is uniquely defined on all of $\mathbf{K}(T)$.

Accordingly, let $\mathbf{x} \in \mathbf{K}(T)$ and suppose $\mathbf{x}=\mathbf{x}(s_1, \cdots, s_{n-1}, s_n, j) \in \mathbf{D}(j, T)$ and the unique control function for $\mathbf{x}$ has $\mathbf{p}(j, \pi)=\langle \mathbf{u}^1, \mathbf{u}^2, \cdots, \mathbf{u}^n \rangle$ (or a subpolicy of $\mathbf{p}(j, \pi)$) as its switching policy and switches consecutively at times $s_{\pi(1)}, \cdots, s_{\pi(n-1)}$ (or a subset of these times). Then define

$$
\begin{aligned}
\mathbf{F}[\mathbf{x}(s_1, \cdots, s_n, j)] &= \mathbf{u}^1 \quad \text{if } 0 < s_{\pi(1)}, \\
&= \mathbf{u}^2 \quad \text{if } 0 = s_{\pi(1)} < s_{\pi(2)}, \\
&= \mathbf{u}^3 \quad \text{if } 0 = s_{\pi(1)} = s_{\pi(2)} < s_{\pi(3)}, \\
&\vdots \\
&= \mathbf{u}^n \quad \text{if } 0 = s_{\pi(1)} = \cdots = s_{\pi(n-1)} < s_n \leqq T, \\
&= \mathbf{0} \quad \text{if } 0 = s_{\pi(1)} = \cdots = s_{\pi(n-1)} = s_n.
\end{aligned} \tag{7.1}
$$

PROPOSITION 7.2. *The time-optimal feedback function,* $\mathbf{F}$, *satisfies the following*:

(1) $\mathbf{F}$ *is piecewise constant and*:

(a) $\mathbf{F}(\mathbf{x})=\mathbf{u}(1, j)=(\boldsymbol{\delta}(j+1), (-1)^{n-1-j}\boldsymbol{\delta}(j))$ *if* $\mathbf{x}$ *lies in the interior of* $\mathbf{D}(j, T)$ *or in the relative interior of an* $(n-1)$*-cell belonging to either of the* $(n-1)$*-dimensional invariant cell complexes* $\mathbf{D}(j|s_j=s_n)$ *and* $\mathbf{D}(j|s_{n-1}=s_n)$.

(b) $\mathbf{F}(\mathbf{x})=(-\boldsymbol{\delta}(j+1), (-1)^{n-j-1}\boldsymbol{\delta}(j))$ *if* $\mathbf{x}$ *lies in the relative interior of an* $(n-1)$*-cell belonging to the* $(n-1)$*-dimensional attracting cell complex* $\mathbf{D}(j|s_1=0)$.

(c) $\mathbf{F}(\mathbf{x})=(\boldsymbol{\delta}(j+1), (-1)^{n-j}\boldsymbol{\delta}(j))$ *if* $\mathbf{x}$ *lies in the relative interior of an* $(n-1)$*-cell belonging to the* $(n-1)$*-dimensional attracting cell complex* $\mathbf{D}(j|s_{j+1}=0)$.

(2) *The time-optimal trajectories of* (1.1) *on* $\mathbf{K}(T)$ *are solutions of*

$$
\dot{\mathbf{x}}=\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{F}(\mathbf{x}), \qquad \mathbf{x}(0)=\mathbf{x}^0 \in \mathbf{K}(T). \tag{7.3}
$$

(3) *If* $\mathbf{x}$ *lies in the relative interior of the* $(n-\alpha)$*-cell* $\mathbf{D}(j, \pi, \alpha)$ (*defined in* (5.3)) *for* $\alpha=1, \cdots, n-1$, *the vector* $\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{F}(\mathbf{x})$ lies in the tangent space of $\mathbf{D}(j, \pi, \alpha)$ *at* $\mathbf{x}$ *and the vector field* $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{F}(\mathbf{x})$ *is transversal to the* $(n-\alpha-1)$*-cell,* $\mathbf{D}(j, \pi, \alpha+1)$.

Proof. The proof of these assertions follows directly from the definition of $\mathbf{F}$, (5.1)-(5.3), and the fact that the "jump" in the tangent vector to an optimal trajectory traversing $\mathbf{D}(j, \pi, \alpha)$ at a point of $\mathbf{D}(j, \pi, \alpha+1)$ is (from (5.2))

$$
\begin{aligned}
\dot{\varphi}(s_{\pi(\alpha+1)}+0)-\dot{\varphi}(s_{\pi(\alpha+1)}-0) &= \mathbf{B}(\mathbf{u}^{\alpha+1}-\mathbf{u}^{\alpha}) \\
&= \pm\mathbf{b}^1 \quad \text{or} \quad \pm\mathbf{b}^2,
\end{aligned}
$$

depending on the coordinate in which the $\alpha+1$ switch occurs. □

COROLLARY 7.4. *The collection of cells composed of the following*:

(1) *The interiors,* $\pm\mathbf{D}(j, T)^0$, $j=0, \cdots, n-1$, *of the policy complexes*;

(2) *The cells* $\mathbf{D}(j, \pi, \alpha)$, $\alpha=1, \cdots, n-1$, (*see* (5.3)) *for all* j*-permissible permutations* π *and for all* $j=0, \cdots, n-1$;

(3) *The cell* $\{\mathbf{0}\}$

with the time-optimal feedback function, $\mathbf{F}$, *of* (7.1) *form a* **regular synthesis** *on* $\mathbf{K}(T)$ *in which every cell is of type* I (*see* [5], [27]).

Proof. This result follows directly from (7.2) and (5.3). □

Since the cells of (7.4) are fully constructable using the maps of (4.5), this result provides the first explicit construction of a regular synthesis for systems of the generality considered here.

Equation (7.3) provides an example of a differential equation "with discontinuous right-hand side" as studied by Filippov [6], Hermes [12], and Hájek [10], [11]. For our purposes the most useful concept of "solution" for such systems is the one proposed by Filippov.

DEFINITION 7.5. Let $\mathbf{f}$ be suitably defined on a subset of $\mathbb{R}^n$ (see [6], [10], and [11]), then an absolutely continuous function φ is said to be a solution to

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \qquad \mathbf{x}(0) = \mathbf{x}^0$$

on the interval $0 \leqq t \leqq T$, in **the sense of Filippov** (an $\mathcal{F}$-solution) if $\varphi(0) = \mathbf{x}^0$ and $\dot{\varphi}(t) \in \mathcal{F}(\mathbf{f}, \mathbf{x}(t))$ almost everywhere on $[0, T]$, where $\mathcal{F}(\mathbf{f}, \mathbf{x})$ is the closed convex set

$$\mathcal{F}(\mathbf{f}, \mathbf{x}) \equiv \bigcap_{\delta>0} \bigcap_{\mu(E)=0} \overline{\text{co}}\,[\mathbf{f}(B(\delta, \mathbf{x}) \backslash E)], \tag{7.6}$$

where $B(\delta, \mathbf{x})$ is the intersection of a δ-ball about $\mathbf{x}$ with the domain of $\mathbf{f}$, μ is n-dimensional Lebesgue measure and $\overline{\text{co}}$ denotes the closed convex hull of its argument set.

Clearly, if $\mathbf{f}$ is continuous at $\mathbf{x}$, $\mathcal{F}(\mathbf{f}, \mathbf{x}) = \{\mathbf{f}(\mathbf{x})\}$. At the other extreme, values of $\mathbf{f}$ taken on near $\mathbf{x}$ only on sets of measure zero play no role in the determination of $\mathcal{F}(\mathbf{f}, \mathbf{x})$ or for $\mathcal{F}$-solutions through $\mathbf{x}$.

DEFINITION 7.7. A feedback function $\mathbf{F}: \mathbf{K} \mapsto \mathscr{C}^2$ is **of Filippov type at** $\mathbf{x} \in \mathbf{K}$ **if** $\mathbf{F}(\mathbf{x}) \in \mathcal{F}(\mathbf{F}, \mathbf{x})$. **F is realizable** on a subset S of $\mathbf{K}$ if it is of Filippov type at each point of S.

Remark 7.8. A practical feedback function should not require "infinitely precise" measurements for its implementation. The determination of function values taken only on sets of measure zero would require such extreme measurement precision. This observation motivates the definition of "realizable" feedback function.

The following proposition generalizes the results concerning the measurement requirements of the feedback function discussed earlier for the $n = 2$ and $n = 3$ cases pictured in Figs. 1-3.

PROPOSITION 7.9. *Let* $\mathbf{F}$ *be the time-optimal feedback function defined in* (7.1), $\mathbf{F}$ *is of Filippov type at each point* $\mathbf{x}$ *of* $\mathbf{K}(T)$ *that lies*

(1) *In the interior of a policy cell complex* $\pm\mathbf{D}(j, T)$;

(2) *In the relative interior of an* $(n-1)$*-cell of the invariant cell complex* $\mathbf{D}(j|s_j = s_n) \cup \mathbf{D}(j|s_{n-1} = s_n) \cup (-\mathbf{D}(j|s_j = s_n)) \cup (-\mathbf{D}(j|s_{n-1} = s_n))$;

(3) *In the relative interior of an* $(n-1)$*-cell of the attracting cell complex* $\pm\mathbf{D}(j|s_1 = 0)$ *if and only if* $\gamma(j) = 1$;

(4) *In the relative interior of an* $(n-1)$*-cell of the attracting cell complex* $\pm\mathbf{D}(j|s_{j+1} = 0)$ *if and only if* $\gamma(j+1) = 1$

for $j = 0, \cdots, n-1$.

Proof. If $\mathbf{x}^0$ lies in the interior or in the relative interior of an $(n-1)$-cell of an invariant cell complex of $\mathbf{D}(j, T)$, Proposition 7.2(1)(a) shows that $\mathbf{F}(\mathbf{x}^0) = \mathbf{u}(1, j)$. In either case, this value is assumed in a neighborhood or a half-neighborhood of $\mathbf{x}^0$ and therefore is assumed on a set of positive measure in every $B(\delta, \mathbf{x}^0)$. This proves assertions (1) and (2) of Proposition 7.9.

If $\mathbf{x}^0$ lies in the relative interior of an attracting $(n-1)$-cell of $\mathbf{D}(j|s_1 = 0)$, then, from the structure Lemma 6.1, it also lies in an invariant cell complex, $-\boldsymbol{\delta}(j) \cdot \boldsymbol{\delta}(j-1) \cdot \mathbf{D}(j-1|s_{n-1} = s_n)$, or an attracting cell complex, $\boldsymbol{\delta}(j) \cdot \boldsymbol{\delta}(j-1) \cdot \mathbf{D}(j-1|s_{j+1} = 0)$, of

$\pm\mathbf{D}(j-1, T)$, when $\boldsymbol{\gamma}(j)=+1$ or -1, respectively. In the first case, $\mathbf{F}$ is realizable at $\mathbf{x}^0$ by the previously proven assertion (2). In the second case, the value $\mathbf{F}(\mathbf{x}^0)$ differs from $\mathbf{u}(1, j)$ in the first coordinate and from $\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{u}(1, j-1)$, the value assumed by $\mathbf{F}$ in the interior of $\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1, T)$, in the second coordinate. This implies that this latter value is, in fact, $-\mathbf{u}(1, j)$. Consequently, in any sufficiently small ball, $B(\delta, \mathbf{x}^0)$, $\mathbf{F}$ takes on only three values: $\mathbf{u}(1, j)$ on points of the interior of $\mathbf{D}(j, T)$; $-\mathbf{u}(1, j)$ on points of the interior of $\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1, T)$; and $\mathbf{F}(\mathbf{x}^0)$ on points of the $(n-1)$-cell $\mathbf{D}(j|s_1=0)$. Since the set of points $\mathbf{x}$ where $\mathbf{F}(\mathbf{x})=\mathbf{F}(\mathbf{x}^0)$ is a set of measure zero and $\mathbf{F}(\mathbf{x}^0)$ is not a convex combination of $\pm\mathbf{u}(1, j)$, $\mathbf{F}$ is not realizable at $\mathbf{x}^0$. This proves (3), and (4) follows from similar arguments. □

8. The time-optimal switching surfaces. The time-optimal feedback function $\mathbf{F}$ is conceptually simple. Its two components F_1 and F_2 assume only the values ±1. As a result the controllable set $\mathbf{K}$ is decomposed, for each $i=1, 2$, into two sets $F_i^{-1}(\pm1)$. The ith switching surface, $\boldsymbol{\Omega}_i$, forms the boundary of these two sets. Moroz [25] has shown that $\boldsymbol{\Omega}_i$ has a one-to-one projection on the subspace orthogonal to $\mathbf{b}^i$.

While examples show that switching surfaces can, in general, be very complicated, the analysis presented above permits a complete description and study of the switching surfaces within the controllable set $\mathbf{K}(T)$, where the results developed in the previous sections apply. The switching surfaces are, in fact, precisely the attracting cell complexes described in (5.4).

PROPOSITION 8.1. *Let system* (1.1) *be strictly normal and* $r=2$. *Then*

$$\text{(a)}\qquad \boldsymbol{\Omega}_1(T)\equiv\boldsymbol{\Omega}_1\cap\mathbf{K}(T)=\bigcup_{j=1}^{n-1}[-\mathbf{D}(j|s_1=0)\cup\mathbf{D}(j|s_1=0)],$$

$$\text{(b)}\qquad \boldsymbol{\Omega}_2(T)\equiv\boldsymbol{\Omega}_2\cap\mathbf{K}(T)=\bigcup_{j=0}^{n-2}[-\mathbf{D}(j|s_{j+1}=0)\cup\mathbf{D}(j|s_{j+1}=0)].$$

Proof. The proof follows by construction and Proposition 7.2. □

PROPOSITION 8.2. *For each* $j=1, 2, \cdots, n-1$, *the* $(n-1)$*-dimensional cell complex* $\mathbf{D}(j|s_1=0)$ *lies in* $\boldsymbol{\Omega}_1(T)\cap\Omega_2(T)$ *if and only if* $\boldsymbol{\gamma}(j)=-1$; *and, for each* $j=0, 1, \cdots, n-2$, *the* $(n-1)$*-dimensional cell complex* $\mathbf{D}(j|s_{j+1}=0)$ *lies in* $\boldsymbol{\Omega}_1(T)\cap\boldsymbol{\Omega}_2(T)$ *if and only if* $\boldsymbol{\gamma}(j+1)=-1$.

Proof. This result follows directly from Lemma 6.1 (the structure lemma) and Proposition 8.1 (e.g., if $\boldsymbol{\gamma}(j)=-1$, $\mathbf{D}(j|s_1=0)$, which lies in $\boldsymbol{\Omega}_1$, is equal to $\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_j=0)$, which lies in $\boldsymbol{\Omega}_2$). □

This description of the switching surfaces coupled with the analytic mappings of (4.5) permit a complete analysis and construction of the switching surfaces for small response times. Such results formed the basis for construction of closed-loop time-optimal controllers for the third- and fourth-order systems reported in [17] and [18]. The switching policies describing the switching surfaces of our example systems are the following.

Example A.

$\boldsymbol{\Omega}_1$: $\langle1, 4\rangle$, $\langle2, 1\rangle$, $\langle3, 2\rangle$, and $\langle4, 3\rangle$,

$\boldsymbol{\Omega}$: $\langle4, 1\rangle$, $\langle1, 2\rangle$, $\langle2, 3\rangle$, and $\langle3, 4\rangle$.

Example B.

$\boldsymbol{\Omega}_1=\boldsymbol{\Omega}_2$: $\langle4, 1\rangle$, $\langle2, 3\rangle$, $\langle2, 1\rangle$, and $\langle4, 3\rangle$.

Example C.

$\mathbf{\Omega}_1$: $\langle 2,3,2\rangle, \langle 1,2,3\rangle, \langle 1,4,3\rangle, \langle 2,1,2\rangle, \langle 4,1,4\rangle, \langle 3,4,1\rangle, \langle 3,2,1\rangle$, and $\langle 4,3,4\rangle$,

$\mathbf{\Omega}_2$: $\langle 1,4,1\rangle, \langle 2,1,4\rangle, \langle 2,3,4\rangle, \langle 1,2,1\rangle, \langle 3,2,3\rangle, \langle 4,3,2\rangle, \langle 4,1,2\rangle$, and $\langle 3,4,3\rangle$.

9. Canonical structures: proof of Theorem 1.8. In this section we present the proof of Theorem 1.8, which, for convenience, we restate below. Recall that

$$\gamma(j) = \gamma(j; \mathbf{A}, \mathbf{B}) \equiv \delta(j-1) \cdot \delta(j+1), \qquad j = 1, \cdots, n-1,$$

$$N(\mathbf{A}, \mathbf{B}) = \sum_{j=1}^{n-1} (1 - \gamma(j; \mathbf{A}, \mathbf{B}))/2,$$

$$M(\mathbf{A}, \mathbf{B}) \equiv \sum_{j=1}^{n-1} (1 - \gamma(j; \mathbf{A}, \mathbf{B}))\binom{n-2}{j-1}.$$

THEOREM 1.8. *Let system* (1.1) *be strictly normal, let* $r = 2$, *and let* T *be sufficiently small and positive. Then the following apply*:

(a) $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$ *has connectivity* $2(n - N(\mathbf{A}, \mathbf{B}))$;

(b) $\mathbf{\Omega}_1(T) \cap \mathbf{\Omega}_2(T)$ *is the union of* $M(\mathbf{A}, \mathbf{B})$ $(n-1)$*-dimensional cells*;

(c) $\mathbf{\Omega}_1(T) = \mathbf{\Omega}_2(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) = n-1$, (*that is,* $\gamma(1; \mathbf{A}, \mathbf{B}) = \cdots = \gamma(n-1; \mathbf{A}, \mathbf{B}) = -1$);

(d) *The time-optimal feedback function is realizable on* $\mathbf{K}(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) = 0$ (*that is,* $\gamma(1; \mathbf{A}, \mathbf{B}) = \cdots = \gamma(n-1; \mathbf{A}, \mathbf{B}) = 1$); *and*

(e) *The feedback function is of Filippov type on the relative interior of some* $(n-1)$*-dimensional cell of* $\mathbf{\Omega}(T)$ *if and only if* $N(\mathbf{A}, \mathbf{B}) \leqq n-2$.

Proof. (a) We have seen that $\mathbf{K}(T)$ is the union of the cell complexes $\pm\mathbf{D}(j, T)$, for $j = 0, \cdots, n-1$. From Lemma 6.1 it is clear that $\mathbf{D}(j, T)$ adjoins $\pm\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$ and $\pm\delta(j) \cdot \delta(j+1) \cdot \mathbf{D}(j+1, T)$ (when $0 < j < n-1$, if $j = 0$ or $n-1$, only one such relationship exists). Its intersection with the former pair consists of the cell complexes $\mathbf{D}(j|s_1 = 0)$ and $\mathbf{D}(j|s_j = s_n)$. If $\gamma(j) = 1$, these complexes lie in $\mathbf{\Omega}_1$ and $\mathbf{\Omega}_2$, respectively (from (8.2)); each intersection is of the attracting/invariant type; and $\mathbf{D}(j, T)$ and $\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$ lie in distinct components of $\mathbf{K}(T)\backslash\Omega(t)$ as, of course, is also true of $-\mathbf{D}(j, T)$ and $-\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$. Similarly, if $\gamma(j+1) = 1$, the intersections of $\mathbf{D}(j, T)$ and $-\delta(j) \cdot \delta(j+1) \cdot \mathbf{D}(j+1, T)$ are of attracting/invariant type and the two n-complexes also lie in distinct components. Thus, if all $\gamma(k) = 1$ and therefore $N(\mathbf{A}, \mathbf{B}) = 0$, each of the $2n$ policy complexes lies in a distinct component of $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$. Consequently, $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$ has connectivity $2n$ when $N(\mathbf{A}, \mathbf{B}) = 0$.

On the other hand, if $\gamma(j) = -1$, the intersection of $\mathbf{D}(j, T)$ with $\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$ is of attracting/attracting type and its intersection with $-\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$ is of invariant/invariant type. The first implies that $\mathbf{D}(j, T)$ and $\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$ are separated by $(n-1)$-cells which lie in $\mathbf{\Omega}_1 \cap \mathbf{\Omega}_2$ and therefore lie in distinct components of $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$. The invariant/invariant intersection, however, obviously implies that the feedback function is continuous on the interior of $\mathbf{D}(j, T) \cup (-\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T))$ and that these two policy complexes lie in the same component. The same is true of $-\mathbf{D}(j, T)$ and $\delta(j) \cdot \delta(j-1) \cdot \mathbf{D}(j-1, T)$. Thus, for each $\gamma(j) = -1$, the connectivity of $\mathbf{K}(T)\backslash\mathbf{\Omega}(T)$ decreases by 2 and $N(\mathbf{A}, \mathbf{B})$ increases by 1. Clearly, this is the relationship presented in assertion (a).

(b) If all $\gamma(j) = 1$, all intersections of the policy complexes are of the attracting/invariant type, $\mathbf{\Omega}_1(T) \cap \mathbf{\Omega}_2(T)$ contains no cells of dimension $n-1$ and $M(\mathbf{A}, \mathbf{B}) = 0$. However, if $\gamma(j) = -1$, the attracting/attracting intersection of $\mathbf{D}(j, T)$ with

$\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1,T)$ in the $(n-1)$-dimensional cell complex $\mathbf{D}(j|s_1=0)=\boldsymbol{\delta}(j)\cdot\boldsymbol{\delta}(j-1)\cdot\mathbf{D}(j-1|s_j=0)$ implies that the

$$\binom{n-2}{j-1}$$

$(n-1)$-dimensional cells of $\mathbf{D}(j|s_1=0)$ lie in both switching surfaces. Then, by symmetry, the minimum number of $(n-1)$-cells in $\boldsymbol{\Omega}_1(T)\cap\boldsymbol{\Omega}_2(T)$ increases by

$$2\cdot\binom{n-2}{j-1}$$

if $\boldsymbol{\gamma}(j)=-1$. This proves (b).

(c) If all $\boldsymbol{\gamma}(j)=-1$, all policy complexes intersect in attracting/attracting combinations, each attracting cell lies in the intersection of the two switching surfaces and, since lower-dimensional cells all lie in each surface, $\boldsymbol{\Omega}_1(T)=\boldsymbol{\Omega}_2(T)$.

(d) If $N(\mathbf{A},\mathbf{B})$ is positive, there exists at least one $\boldsymbol{\gamma}(j)=-1$ and at least two $(n-1)$-cells in $\boldsymbol{\Omega}_1(T)\cap\boldsymbol{\Omega}_2(T)$. In this case the time-optimal feedback function $\mathbf{F}$ is not realizable on these cells by Proposition 7.9(4). However, if $N(\mathbf{A},\mathbf{B})=0$, all policy complex intersections are of the attracting/invariant type. As a result each time-optimal attracting subpolicy of one complex is simultaneously an invariant subpolicy of an adjacent policy complex. It is this property that permits us to show that $\mathbf{F}$ is of Filippov type at each point of $\mathbf{K}(T)$ and, therefore, realizable.

Now let $\mathbf{x}^0\in\mathbf{K}(T)$. If $\mathbf{x}^0$ lies in the interior of a policy complex or in the relative interior of an $(n-1)$-dimensional boundary complex, then $\mathbf{F}$ is of Filippov type at $\mathbf{x}^0$ by (7.9). Therefore, suppose $\mathbf{x}^0$ lies in some lower-dimensional boundary cell of $\mathbf{D}(j,\pi,T)$ and that the time-optimal control function for $\mathbf{x}^0$ has a switching policy that is a subpolicy of $\mathbf{p}(j,\pi)=\langle\mathbf{u}^1,\mathbf{u}^2,\cdots,\mathbf{u}^n\rangle$. Specifically, suppose $\mathbf{F}(\mathbf{x}^0)=\mathbf{u}^k$ and that the policy describing the control function for $\mathbf{x}^0$ is (a subpolicy of)$\langle\mathbf{u}^k,\mathbf{u}^{k+1},\cdots,\mathbf{u}^n\rangle$. Thus $\mathbf{x}^0$ lies in the $(n-k+1)$-dimensional cell $\mathbf{D}(j,\pi,k-1)$ (see (5.3)) corresponding to this control policy.

To show that $\mathbf{F}$ is of Filippov type at $\mathbf{x}^0$, it is sufficient to show that $\mathbf{F}$ takes on the value $\mathbf{u}^k$ on a set of positive measure inside any neighborhood of $\mathbf{x}^0$. To do this it is sufficient to show that $\mathbf{x}^0$ lies in a boundary cell of some policy complex, say $\mathbf{D}(j',T)$, with $\mathbf{F}(\mathbf{x})=\mathbf{u}(1,j')=\mathbf{u}^k$ on points, $\mathbf{x}$, in its interior.

The attracting subpolicy $\langle\mathbf{u}^2,\cdots,\mathbf{u}^n\rangle$ *of* $\mathbf{p}(j,\pi)$ is also an invariant policy of an n-policy $\langle\mathbf{u}^2,\cdots,\mathbf{u}^n,\mathbf{w}^1\rangle$ of an adjacent policy complex. In turn, the attracting subpolicy of this n-policy is an invariant subpolicy of another n-policy $\langle\mathbf{u}^3,\cdots,\mathbf{u}^n,\mathbf{w}^1,\mathbf{w}^2\rangle$. Clearly this process continues until we identify a time-optimal policy $\mathbf{p}(j',\pi')=\langle\mathbf{u}^k,\cdots,\mathbf{u}^n,\mathbf{w}^1,\cdots,\mathbf{w}^{k-1}\rangle$, where j' denotes the number of switches in the first control coordinate implied by the policy. Because the control policy of $\mathbf{x}^0$ is a subpolicy of $\mathbf{p}(j',\pi')$, $\mathbf{x}^0$ is a point of the boundary of $\mathbf{D}(j',T)$. It is clear that points of the boundary lie in the closure of the interior of $\mathbf{D}(j',T)$. Hence there exist interior points arbitrarily close to $\mathbf{x}^0$. Since each such interior point has a neighborhood in which $\mathbf{F}\equiv\mathbf{u}^k$, $\mathbf{F}$ is of Filippov type at $\mathbf{x}^0$ and (d) is proven.

(e) If $N(\mathbf{A},\mathbf{B})=n-1$ then $\boldsymbol{\Omega}_1(T)=\boldsymbol{\Omega}_2(T)$ and $\mathbf{F}$ is not realizable at any point of the relative interior of an $(n-1)$-dimensional cell of $\boldsymbol{\Omega}(T)$. On the other hand, if $N(\mathbf{A},\mathbf{B})\leqq n-2$, there exists at least one "attracting/invariant" intersection of $(n-1)$-cells. As has been previously shown, $\mathbf{F}$ is of Filippov type on the relative interior of each such cell. This completes the proof of Theorem 1.8. □

COROLLARY 9.1. *Under the hypotheses of Theorem* 1.8 *all time-optimal trajectories on* $\mathbf{K}(T)$ *are Filippov trajectories if and only if* $\boldsymbol{\gamma}(1;\mathbf{A},\mathbf{B})=\cdots=\boldsymbol{\gamma}(n-1;\mathbf{A},\mathbf{B})=1$.

Proof. A time-optimal trajectory through a point $\mathbf{x}^0$ is a Filippov trajectory if and only if $\mathbf{F}$ is of Filippov type at $\mathbf{x}^0$. □

Theorem 1.8 shows that the time-optimal flow and switching surface structure near $\mathbf{0}$ is completely determined, qualitatively, by the structure constants $\gamma(1), \cdots, \gamma(n-1)$. These integers, unlike the determinants $\mathbf{d}(0), \cdots, \mathbf{d}(n)$ are invariant under linear transformations and provide a canonical classification of strictly normal systems. Since the system is strictly normal it is possible to use the columns of any $\mathbf{d}(j)$ as a basis to transform the system into its equivalent Luenberger canonical form $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ [15]. That is, if $0 < j < n$, and $\mathbf{T}$ denotes the matrix corresponding to the determinant $\mathbf{d}(j)$,

$$\bar{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = [\mathbf{e}^2, \cdots, \mathbf{e}^j, \mathbf{a}, \mathbf{e}^{j+2}, \cdots, \mathbf{e}^n, \mathbf{c}], \qquad \bar{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} = [\mathbf{e}^1, \mathbf{e}^{j+1}]$$

where $\mathbf{e}^k$ denotes the kth column of the $n \times n$ identity matrix, $\mathbf{a} = \mathbf{T}^{-1}\mathbf{A}^j\mathbf{b}^1$ and $\mathbf{c} = \mathbf{T}^{-1}\mathbf{A}^{n-j}\mathbf{b}^2$.

With the system in canonical form, it can be shown that for the low-dimensional systems of primary interest, $n \leqq 6, \gamma(1), \cdots, \gamma(n-1)$ constitute a system of independent invariants and provide a complete canonical categorization of linear time-optimal control systems. This result, however, remains to be proven for general n.

10. L^*-systems. Olsder [26] introduced a class of systems he called L^*-systems. In the case of two-dimensional controls of interest here, system (1.1) is L^* if for $n = 2k$, $\mathbf{d}(k-1)$, $\mathbf{d}(k)$, and $\mathbf{d}(k+1)$ are nonzero, or if for $n = 2k+1$, $\mathbf{d}(k)$ and $\mathbf{d}(k+1)$ are nonzero.

For such systems Olsder shows that, for almost all $\mathbf{x}$ of the unit sphere in $\mathbb{R}^n$, there exists an $\varepsilon_{\mathbf{x}} > 0$ such that $\varepsilon_{\mathbf{x}}\mathbf{x}$ is time-optimally controlled to $\mathbf{0}$ by a unique control function having k switches in each coordinate when $n = 2k+1$ and k switches in one coordinate and $k-1$ switches in the other coordinate when $n = 2k$. Furthermore, the switching times and response times are analytic functions of a power of $\varepsilon_{\mathbf{x}}$.

In terms of the results developed in this paper, Olsder's work implies that in the $n = 2k+1$ case, for example, almost every ray from the origin intersects $-\mathbf{D}(k, T) \cup \mathbf{D}(k, T)$ in an interval containing $\mathbf{0}$ in its interior. Interpreted geometrically, this implies that the switching surfaces separating these complexes from the others meet tangentially at the origin. Furthermore, as $T \to 0$ the relative contribution of all other complexes $(\pm\mathbf{D}(j, T), j \neq k)$ to $\mathbf{K}(T)$ becomes negligible.

The two complexes $\pm\mathbf{D}(k, T)$ lie "above" and "below" both switching surfaces and contain the major portion of $\mathbf{K}(T)$. This is illustrated in Fig. 3 and evident in Examples A and B and the general analysis of [17], [20].

11. Stability with respect to measurement. Hermes [12] has introduced the concept of "stability with respect to measurement" to characterize those feedback systems

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{F}(\mathbf{x}), \tag{11.1}$$

that are tolerant of measurement error. Hermes' original formulation has been shown by Hájek [10], [11] to be equivalent to "stability with respect to inner perturbations" (an *inner perturbation* of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ is of the form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x} + \mathbf{p}(\mathbf{x}))$ where $\mathbf{p}$ is essentially bounded). This formulation has proved to be more tractable than the original used by Hermes.

There are necessary conditions for measurement stability and sufficient conditions for measurement stability, but necessary *and* sufficient conditions only for the case of scalar controls, arbitrary n (see [11]), two-dimensional controls with $n = 2$ (see [3], [4]), and two-dimensional controls with $n = 3$ (see [21]). Furthermore, only [3] and

[4] deal with the global case; the other results consider only "local measurement stability" in a neighborhood of the origin.

PROPOSITION 11.2. *A necessary condition for a closed-loop time-optimal control system of the form* (11.1) *to be locally stable with respect to measurement is that* **F** *be realizable on some* $\mathbf{K}(T)$, $T>0$.

Proof. In his original paper [12] Hermes showed that if a system is measurement stable, every Carathéodory solution of (11.1) is also a Filippov solution ($\mathcal{F}$-solution). If **F** is not realizable, there exist Carathéodory solutions on every $\mathbf{K}(T)$ that are not $\mathcal{F}$-solutions. On the other hand, if **F** is realizable, every Carathéodory solution is an $\mathcal{F}$-solution. □

COROLLARY 11.3. *Let system* (1.1) *be strictly normal and* $r=2$; *then the condition*

$$\boldsymbol{\gamma}(1;\mathbf{A},\mathbf{B})=\cdots=\boldsymbol{\gamma}(n-1;\mathbf{A},\mathbf{B})=1$$

is necessary for the existence of a locally measurement stable time-optimal feedback function.

Proof. The proof follows directly from Theorem 1.8(d). □

PROPOSITION 11.4. *A closed-loop time-optimal control system of the form* (11.1) *with r-dimensional controls,* $1\leqq r\leqq n$, *and a realizable time-optimal feedback function on some* $\mathbf{K}(T)$, $T>0$, *is locally measurement stable if and only if Filippov solutions to* (11.1) *are unique for all* $\mathbf{x}^0\in\mathbf{K}(T)$.

Proof. If **F** is realizable on $\mathbf{K}(T)$, the classes of Filippov solutions, Krasovsky solutions, and Hermes solutions to (11.1) (see [10], [11]) coincide. The result then follows from [11, Lemma 9.1], which states that (11.1) is stable with respect to measurement if and only if Hermes solutions are unique. □

Unfortunately, existence of a realizable feedback function is not sufficient to imply uniqueness of Filippov solutions. The following result from [21] exhibits an additional requirement.

THEOREM 11.5. *A strictly normal system of the form* (1.1) *with* $n=3$ *and* $r=2$ *is locally stable with respect to measurement if and only if*

(a) $\boldsymbol{\gamma}(1;\mathbf{A},\mathbf{B})=\boldsymbol{\gamma}(2;\mathbf{A},\mathbf{B})=1$, *and*

(b) $\det[\mathbf{b}^1,\mathbf{b}^2,\boldsymbol{\delta}(2)\mathbf{E}^1(t)+\boldsymbol{\delta}(1)\mathbf{E}^2(t)]\not\equiv 0$.

Remark 11.6. In [21] the determinants $\mathbf{d}(j)$ are indexed by the number of $\mathbf{b}^2$-based columns rather than $\mathbf{b}^1$-based columns. As a consequence, the roles of $\boldsymbol{\delta}(1)$ and $\boldsymbol{\delta}(2)$ are reversed in the original paper.

The system of Example A is locally measurement stable.

In general, switching surfaces $\boldsymbol{\Omega}(T)$ will contain Boltyanskii "cells of the second kind" (see [2], [5]) for sufficiently large T. However, it is unlikely that such cells could introduce measurement instability. Indeed, with the proper cellular decomposition these cells should occur as attracting sets for some n-cell and "start-points" (see [1]) for another n-cell complex. If this turns out to be an accurate view of the switching surface structure, then Filippov solutions would be locally unique in a neighborhood of cells of the second kind. This analysis implies the following conjecture.

Conjecture 11.7. A time-optimal feedback system of the form (11.1) with r-dimensional controls, $1\leqq r\leqq n$, is stable with respect to measurement if and only if it is locally stable with respect to measurement.

12. Analysis of more general systems. The techniques employed here to analyze strictly normal two-input systems may be extended to general minimally controllable r-input systems. This work is in progress and will be reported later. In addition, the total information concerning the local regular synthesis gained from this analysis sheds

considerable light on the total switching surface structure for large T. It is expected that this information will permit a resolution of (11.7).

13. Summary. This paper has extended the analytical techniques of [20] to the general nth-order strictly normal two-input system. In this extension:

(1) A cellular decomposition of $\mathbf{K}(T)$, for sufficiently small T, has been described and shown to provide the first explicit construction of a local regular synthesis for multi-input systems of arbitrary order.

(2) A system of linear invariants has been identified and been shown to completely characterize the time-optimal flow and switching surface structure for small response times.

(3) New necessary, and necessary and sufficient, conditions for local measurement stability have been proven.

REFERENCES

[1] N. P. BHATIA AND O. HÁJEK, *Local Semi-Dynamical Systems*, Springer-Verlag, New York, 1969.

[2] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Holt, Reinhart, and Winston, New York, 1971.

[3] P. BRUNOVSKY, *The closed-loop time-optimal control.* I: *Optimality*, SIAM J. Control Optim., 12 (1974), pp. 624–634.

[4] ———, *The closed-loop time-optimal control.* II: *Stability*, SIAM J. Control Optim., 14 (1976), pp. 156–162.

[5] ———, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81–100.

[6] A. F. FILIPPOV, *Differential equations with discontinuous right-hand sides*, Trans. Amer. Math. Soc., 42 (1964), pp. 199–231.

[7] H. FLANDERS, *Differential Forms*, Academic Press, New York, 1963.

[8] M. GREENBURG, *Lectures on Algebraic Topology*, W. A. Benjamin, New York, 1966.

[9] O. HÁJEK, *Terminal manifolds and switching locus*, Math. Systems Theory, 6 (1973), pp. 289–301.

[10] ———, *Discontinuous differential equations*, I, J. Differential Equations, 32 (1979), pp. 149–170.

[11] ———, *Discontinuous differential equations*, II, J. Differential Equations, 32 (1979), pp. 171–185.

[12] H. HERMES, *Discontinuous vector fields and feedback control*, in Differential Equations and Dynamical Systems, J. K. Hale and J. P. LaSalle, eds., Academic Press, New York, 1967, pp. 155–165.

[13] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time-Optimal Control*, Academic Press, New York, 1967.

[14] E. LEE AND L. MARCUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[15] D. G. LUENBERGER, *Canonical forms for linear multi-variable systems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 290–293.

[16] A. T. LUNDELL AND S. WEINGRAM, *The Topology of CW Complexes*, Van Nostrand-Reinhold, New York, 1969.

[17] L. D. MEEKER AND N. PURI, *Closed-loop computerized time-optimal control of multi-variable systems*, Proc. Joint Automatic Control Conference, St. Louis, MO, 1971.

[18] L. D. MEEKER AND G. KRAFT, *Closed-loop time-optimal control of the linearized rendevous problem*, in Proc. 7th Annual Princeton Conference on Information Science and Systems, March 1973, pp. 467–471.

[19] L. D. MEEKER, *The synthesis of discontinuous vector fields with application to time-optimal control*, unpublished manuscript, 1972.

[20] ———, *Time-optimal control for small disturbances*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1095–1099.

[21] ———, *Measurement stability of third-order time-optimal control systems*, J. Differential Equations, 36 (1980), pp. 54–65.

[22] ———, *Closed-loop time-optimal control of a satellite*, in Proc. Joint Automatic Control Conference, Vol. 1, San Francisco, August 1980.

[23] ———, *Bang-bang control and local semi-dynamical systems on CW complexes*, in Dynamical Systems II, A. R. Bednarek and L. Cesari, eds., Academic Press, New York, 1982, pp. 613–619.

[24] ———, *Local time-optimal control of two-input minimally controllable linear systems*, in preparation.

[25] A. MOROZ, *Time-optimal control synthesis problem*, Automat. Remote Control, 1 (1970), pp. 18–28.
[26] G. J. OLSDER, *Time-optimal control of multivariable systems near the origin*, J. Optim. Theory Appl., 16 (1975), pp. 497–517.
[27] H. SUSSMANN, *Regular synthesis for time-optimal control of single-input real analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.
[28] A. WALLACE, *Differential Topology*, W. A. Benjamin, New York, 1968.
[29] D. S. YEUNG, *Time-optimal feedback control*, J. Optim. Theory Appl., 21 (1977), pp. 71–82.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 83-107, January 1989

006

CONSISTENT APPROXIMATIONS OF LINEAR STOCHASTIC MODELS*

ANDREA GOMBANI†

Abstract. This paper considers the problem of approximating a stochastic process $\{y(t)\}$ with state space X. The desired process $\{y_1(t)\}$ has state space X_1, of dimension as small as possible, such that, in mean square norm,

$$\|y(t)-y_1(t)\| \leqq \varepsilon$$

for a given $\varepsilon \geqq 0$. The solution given here has the *inclusion property*, i.e., $X_1 \subset X$ and is *consistent*, that is, it reduces to the problem of finding a minimal realization of $y(t)$ when ε is set equal to zero.

Key words. stochastic realization, approximate model, splitting subspaces, ε-observability

AMS(MOS) subject classifications. 93E12, 93B20

1. Introduction.

In this paper we consider the problem of stochastic model reduction. This consists of finding, for a given process $\{y(t)\}$, a process $\{y_1(t)\}$ admitting a Markovian representation of lowest possible dimension that approximates $\{y(t)\}$ in some norm. This problem has received considerable attention in recent years [5], [7], [13], [14], especially in connection with the Hankel-norm approximation of a spectral factor of y. However, in this paper we investigate a different approach to the problem, one which exploits the geometric framework of stochastic realization theory [9], [10], [11]. The basic idea of this approach is to start with a state space X of y (not necessarily minimal) and build the reduced model *within* this space. This is done by cutting off the parts of X that are almost unobservable *or* almost unconstructible (in a sense to be described below).

There are several reasons for considering this approach. One is that the reduced model will have its state space included in the original one, which means that, in a suitable basis, the new model is a subsystem of the original in the sense explained below. A more important reason is that in this way our algorithm is *consistent*, that is, the procedure we propose solves as a special case the problem of finding a *minimal* realization of $\{y(t)\}$, given a nonminimal one.

The Hankel-norm approximation is generally not consistent; this implies that some care must be taken in the choice of the state space X we want to reduce. (In fact, it is shown in [16] that the best results are achieved with this method when the minimum-phase spectral factor is approximated.) So, for the problem of reducing *any given* state space X, the Hankel-norm is not very suitable.

On the other hand, our procedure might not perform any better than the Hankel-norm method in the finite-dimensional case (even if, in the given example, it yields a much better bound and there seems to be an advantage in infinite dimension).

Since we work with approximation of *stochastic* systems, we will use the mean square norm

$$\|x\| := (Ex^2)^{1/2}.$$

* Received by the editors July 7, 1986; accepted for publication (in revised form) April 8, 1988. This research was conducted while the author was a student at the Royal Institute of Technology, Stockholm, Sweden, and was partially supported by the National Swedish Board of Technical Development under grant 83-3272, and by a fellowship from the Italian National Research Council.

† LADSEB-Consiglio Nazionale delle Ricerche, Corso Stati Uniti 4, 35020 Padova, Italy.

This is the natural norm for this setting, and is unitarily equivalent to the L^2 norm. As we recall, the Hankel-norm approximation in stochastic model reduction is generally used to achieve an upper bound on the L^2-norm [14], [16], and it is this context in which this paper refers to Hankel-norm approximation.

It could be argued that the problem, posed in a deterministic setting, could be better solved with H^∞-methods (with respect, for instance, to robustness) (see, e.g., [15]). However, these methods are not very suitable when a variable phase factor is built into the model. This situation is illustrated by the following example (due to J. C. Willems, and pointed to us by C. I. Byrnes).

Suppose we want to record a symphony. Then the different distances between the instruments and the recording apparatus will introduce a phase shift in the signal. This phase shift is not strongly continuous under the H^∞-norm. Therefore, any attempt to filter out noise, eliminate redundancy, etc., will encounter difficulty, since even slightly shifted versions of the same input will be very distant from each other in H^∞, and will have to be treated as different signals. The L^2 norm does not have this disadvantage.

2. Preliminaries. We review here some basic facts about stochastic realization (we refer the reader to [9]–[11] and references therein for the full story).

Let $y(t)$ be a real-valued discrete-time, stationary, centered, purely nondeterministic (p.n.d.), Gaussian process on the probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$. Consider the Hilbert space generated by the process

$$H = \overline{\text{span}}\,\{y(t);\ t \in \mathbb{Z}\}$$

where the closure is taken with respect to the inner product $\langle x, z\rangle = Exz$, E denoting expected value.

The space H comes naturally endowed with a bilateral shift U, defined by

$$Uy(t) := y(t+1)$$

and extended by linearity to the whole space.

The basic realization problem is the following: find *all* representations of the form

$$x(t+1) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t) \tag{2.1}$$

where $u(t)$ is a white noise process in H (i.e., for each t, $s \in \mathbb{Z}$, $Eu(t)u(s) = \delta_{t,s}$), and A, B, C, D are constant matrices. It is of particular interest to characterize all minimal representations of form (1) (minimal in the sense that the vector $x(t)$ has smallest possible dimension). This problem can be posed as a geometric problem in the Hilbert space H, namely, the problem of finding all subspaces X of H that are *Markovian* and *splitting* for y, in the following sense. A nonzero subspace X of H is said to be *Markovian* if it splits its own past and future, i.e., defining $X^+ = \overline{\text{span}}\,\{U^n X;\ n \geqq 0\}$ and $X^- = \overline{\text{span}}\,\{U^n X;\ n \leqq 0\}$; thus we have that

$$E^{X^-}X^+ = E^X X^+, \tag{2.2a}$$

$$E^{X^+}X^- = E^X X^- \tag{2.2b}$$

with E^A denoting the orthogonal projection on A.

The space X is *splitting* for y if it splits the past and future of y, i.e., if $H^+ = \overline{\text{span}}\,\{y(n);\ n \geqq 0\}$ and $H^- = \text{span}\,\{y(n);\ n < 0\}$ the following holds:

$$E^{H^- \vee X}H^+ = E^X H^+, \tag{2.3a}$$

or equivalently,

$$E^{H^+ \vee X}H^- = E^X H^-. \tag{2.3b}$$

For short, in the sequel we use the concept of *state space* to denote a Markovian splitting subspace (for y).

A state space X for y is *minimal* if any other state space $X_1 \subset X$ is necessarily equal to X. Minimality can be characterized geometrically, introducing the concepts of observability and constructibility. We say that a state space X is *observable* if

$$E^X H^+ = X \tag{2.4a}$$

and *constructible* if

$$E^X H^- = X. \tag{2.4b}$$

It is then consistent with these definitions to say that a subspace $X_1 \subset X$ is *unobservable* if $X_1 \subset (H^+)^\perp$, and that $X_2 \subset X$ is *unconstructible* if $X_2 \subset (H^-)^\perp$.

Markovianness can also be characterized in a more convenient way. We say that a subspace $Z \subset H$ is invariant for the shift U if $UZ \subset Z$ and invariant for the adjoint shift U^* if $U^*Z \subset Z$. We say that $Z_1, Z_2 \subset H$ *intersect perpendicularly* if one of the following equivalent conditions holds:

(2.5a) (i) $(Z_1)^\perp \subset Z_2$ ($\perp$ denoting orthogonal complement in H);

(2.5b) (ii) $E^{Z_1} Z_2 = E^{Z_2} Z_1$.

Then it can be shown [11] that any splitting subspace X can be represented as the intersection of a unique pair $(S, \bar{S})$ of perpendicularly intersecting subspaces of H, i.e.,

$$X = S \cap \bar{S}$$

such that

$$S \supset H^- \tag{2.6a}$$

and

$$\bar{S} \supset H^+. \tag{2.6b}$$

The space X is Markovian if and only if $U^*S \subset S$ (i.e., S is invariant for the backward shift and $U\bar{S} \subset \bar{S}$) and $U\bar{S} \subset \bar{S}$ (i.e., $\bar{S}$ is invariant for the forward shift). *Hence we have characterizations of all state spaces of y in H in terms of invariant perpendicularly intersecting subspaces satisfying* (2.6). This correspondence will be denoted by $X \cong (S, \bar{S})$. It can be shown [11] that the following decomposition holds:

$$H = \bar{S}^\perp \oplus X \oplus S^\perp. \tag{2.7}$$

In particular, $S = X \oplus \bar{S}^\perp$, and $\bar{S} = X \oplus S^\perp$. Now the question is how to get a *minimal* state space from a given $X \cong (S, \bar{S})$. The answer is provided by the following theorem and is the guideline of our model reduction scheme.

THEOREM 2.1 [11]. *Let $X \cong (S, \bar{S})$ be a Markovian splitting subspace. Let $\bar{S}_1 := H^+ \vee S^\perp$ and $S_1 := H^- \vee \bar{S}_1^\perp$. Then $X_1 \cong (S_1, \bar{S}_1)$ is a minimal Markovian splitting subspace such that $X_1 \subset X$.*

The importance of this theorem is that the construction of a minimal Markovian splitting subspace from a nonminimal one is a model reduction problem, in which the Markovian and splitting properties are both preserved.

To analyze this geometry in an efficient way, a functions model is needed. Therefore, as is customary, we shall work in the isomorphic setting of the Hardy spaces H^2 and $\bar{H}^2$.

We recall that $L^2(\mathbb{T})$ is the Hilbert space of square integrable complex-valued function on the unit circle.

Every $f \in L^2(\mathbb{T})$ has a representation (unique)

$$f(e^{i\omega}) = \sum_{n=-\infty}^{\infty} f_n e^{i\omega n}.$$

We write $H^2(\bar{H}^2)$ for the subspace of L^2 whose elements have positive (negative) vanishing Fourier coefficients. The symbol $H^\infty(\bar{H}^\infty)$ denotes the subspace of $H^2(\bar{H}^2)$ whose elements are essentially bounded functions. An *outer function* f in H^2 is an element with the property that $fH^2 := \overline{\text{span}}\{fg;\ g \in H^\infty\} = H^2$, and an *inner function* $h \in H^2$ is an element subject to $|h(e^{i\vartheta})| = 1$ almost everywhere.

The symmetric concepts in $\bar{H}^2$ will be called *conjugate outer* and *conjugate inner.* It can be shown that any function $g \in H^2(\bar{H}^2)$ can be factored as $g = fh$ with f (conjugate) outer and h (conjugate) inner.

Remark. The notation used here is not conventional (the standard mathematical notation being exactly the opposite). This choice has been adopted to be in agreement with most engineering literature.

One of the reasons Hardy spaces have become an indispensable tool in realization theory is illustrated by the next theorem. We recall that an $\mathbb{R}$-valued p.n.d., mean square continuous, stationary process $\{y(t)\}$ has the representation

$$\{y(t)\} = \int_{-\pi}^{\pi} e^{i\omega t}\, d\hat{y}(\omega),$$

$d\hat{y}(\omega)$ being an orthogonal random measure such that there exists a finite, positive, absolutely continuous measure $dF(e^{i\omega})$ with density $\phi(e^{i\omega})$, for which

$$E|d\hat{y}(\omega)|^2 = \phi(e^{i\omega})\, d\omega.$$

We also recall that the spectral measure of a white noise is an orthogonal random measure $d\hat{u}(\omega)$ subject to

$$E|d\hat{u}(\omega)|^2 = d\omega.$$

It can be shown that if $u(t)$ is a white noise in H (i.e., $Eu(t)u(s) = \delta_{ts}$), then the spectral measure defined by

$$u(t) = \int_{-\pi}^{\pi} e^{i\omega t}\, d\hat{u}(\omega) \tag{2.8}$$

induces an isometric isomorphism I_u from H onto L^2 defined via the trigonometric polynomials $p(z)$ as

$$I_u p(u)u(0) := p(e^{i\omega})$$

and extended to the whole of H by continuity. (See [12] for details.)

Let K be an inner function. By $H(K)$ we denote the subspace (invariant for the left shift)

$$H(K) = H^2 \ominus KH^2.$$

Analogously, by $\bar{H}^2(K^*)$ we denote the subspace $\bar{H}^2 \ominus K^*\bar{H}^2$.

THEOREM 2.2 [9]-[11]. *Let $X \cong (S, \bar{S})$ be Markovian splitting. Then there exists a unique pair of white noises $(u, \bar{u})$ such that*

$$I_u S = z^{-1}H^2, \tag{2.9a}$$

$$I_u \bar{S} = K\bar{H}^2, \tag{2.9b}$$

$$I_{\bar{u}} S = z^{-1}K^*H^2, \tag{2.9c}$$

$$I_{\bar{u}} \bar{S} = \bar{H}^2 \tag{2.9d}$$

where K is an inner function in H^2. The process $y(t)$ has the representation

$$y(t) = \int_{-\pi}^{\pi} e^{i\omega t} W(e^{i\omega})\, d\hat{u}(\omega) \tag{2.10a}$$

with $W \in H^2$, i.e., $I_u y(0) = W$. Analogously, $I_{\bar{u}} y(0) = \bar{W} \in \bar{H}^2$ and the following representation holds:

$$y(t) = \int_{-\pi}^{\pi} e^{i\omega t} \bar{W}(e^{i\omega})\, d\hat{\bar{u}}(\omega). \tag{2.10b}$$

Moreover,

$$K = W\bar{W}^{-1}.$$

The correspondence between state spaces X and pairs $(W, \bar{W})$ and $(u, \bar{u})$ is one to one. It will be denoted by $X \cong (W, \bar{W}) \cong (u, \bar{u})$. The functions $W(z)$ and $\bar{W}(z)$ are real for real z. The space X has the representations

$$X = \int \mathscr{X}\, d\hat{u} = \int \bar{\mathscr{X}}\, d\hat{\bar{u}} \tag{2.11}$$

where

$$\mathscr{X} = z^{-1}H(K) = z^{-1}(H^2 \ominus KH^2) \tag{2.12a}$$

and

$$\bar{\mathscr{X}} = \bar{H}^2(K^*) = \bar{H}^2 \ominus K^*\bar{H}^2. \tag{2.12b}$$

The functions W and $\bar{W}$ are called, respectively, *stable* and *strictly unstable spectral factors* of y associated to X. Consequently, in the spectral domain, we represent the pair $(S, \bar{S})$ by the pair $(W, \bar{W})$, or equivalently by $(u, \bar{u})$.

Let X be a Markovian splitting subspace for y. How are the pairs (X^-, X^+) and $(S, \bar{S})$ related? Clearly, $X^- \subset S$, and $X^+ \subset \bar{S}$, but it is not a priori clear if equality holds. The next theorem shows this.

THEOREM 2.3. *Let $X \cong (u, \bar{u}) \cong (S, \bar{S})$ be a state space for y. Then, $X^- = S$ and $X^+ = \bar{S}$.*

Proof. Since it can be shown [11] that $S = H^-(u)$, we need to show only that $u(-1) \in X^-$, because then the inclusion follows from the invariance of X^-. To this end, represent $y(t)$ as in [12]:

$$y(t) = \sum_{n=0}^{\infty} w_n u(-n+t) \tag{2.13}$$

where $\sum_{n=0}^{\infty} w_n z^{-n} = W(z)$, the forward spectral factor associated to X. Let $k > 0$ be the index of the first coefficient in (2.11) which is different from zero. Then

$$E^X y(k-1) = E^S \sum_{n=0}^{\infty} w_n(u)(-n+k-1)$$

$$= \sum_{n=k}^{\infty} w_n u(-n+k-1)$$

because $u(t) \perp S$ for $t \geqq 0$, and $u(t) \in S$ for $t < 0$. Similarly,

$$E^X y(k) = E^S \sum_{n=0}^{\infty} w_n u(-n+k)$$

$$= \sum_{n=k+1}^{\infty} w_n u(-n+k).$$

But $E^X y(k-1)$ and $U^* E^X y(k)$ both belong to X^-. Therefore

$$u(-1) = \frac{1}{w_k}[E^X y(k-1) - U^* E^X y(k)]$$

is also in X^-. The same is true for X^+ (use the representation induced by $\bar{u}$ for $y(t)$ instead of (2.13)).

Let $W \in L^\infty$. Then the Hankel operator with symbol W is

$$H_W := E^{H^{2\perp}} M_{W|H^2}.$$

The singular value $\sigma_k(A)$ of an operator A is defined as

$$\sigma_k(A) = \inf\{\|A - C\|;\ \text{rank } C \leqq u\}.$$

Denote by $\mathscr{R}_k$ the set of strictly proper stable rational functions of degree less than or equal to k. Then there exists the following result due to Adamjan, Arov, and Krein [1]:

$$\min_{f \in \mathscr{R}_k} \|H_W - H_f\| = \sigma_k(H_W).$$

Now let $g \in H^{2\perp}$. It is not difficult to see that

$$\|g\|_H := \|H_g\|$$

defines a norm on $H_2^\perp$ called the Hankel-norm. Then the result of [1] says that the Hankel-norm approximation error of a function W with a function $f \in \mathscr{R}_k$ is exactly $\sigma_k(H_W)$.

Finally, let y be given and let $X \cong (W, \bar{W})$ be splitting for y. By Hankel-norm approximation of y (given X) we mean the Hankel-norm approximation of W. □

3. The problem. This section is devoted to the problem formulation, a discussion of the consistency and inclusion properties, and the illustration of how methods already developed in the literature (such as Hankel-norm approximation) behave with respect to these properties.

We are considering, for the time being, the geometric structure of H, and therefore a coordinate-free approach is appropriate. Hence, given a process $z(t)$, and a Markovian splitting subspace Z for z, it is reasonable to use the word *model* to denote the pair (z, Z). We define the degree of the model to be the dimension of Z.

DEFINITION. We say that the model (y_1, X_1) is a submodel of the model (y, X) if $X \supset X_1$.

This definition is more restrictive than it might seem at first sight, and X_1 cannot be just any subspace of X. In fact, since (y_1, X_1) is a model, X_1 is Markovian, and if X is finite-dimensional, the number of its Markovian subspaces is also finite. Under the assumption $X^- = X_1^-$ this is equivalent to saying that there exists a basis $x(0)$ in X such that the Markov process

$$x(t+1) = Ax(t) + Bu(t)$$

(u is A white noise) can be split into

$$\begin{bmatrix} x_1(t+1) \\ x_2(t+1) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + Bu(t)$$

with x_1 basis for X_1 and $A_{12} = 0$ (which yields Markovianness of the process $x_1(t)$).

We consider two equivalent formulations of the model reduction problem. First define, for $\varepsilon \geqq 0$, the class $\mathscr{Z}_\varepsilon$ to be the class of models (z, Z) satisfying

$$\|z(t) - y(t)\| \leqq \varepsilon \tag{3.1}$$

and, for $k \in \mathbb{N}$, the class $\mathscr{Z}_k$ to be the set of models (z, Z) such that $\dim Z \leqq k$. Finally, the set of admissible models, which will depend on the approximation procedure, will be denoted by $\mathscr{Z}$.

PROBLEM 1. *Given a model* (y, X) *and* $\varepsilon \geqq 0$, *find in* $\mathscr{Z} \cap \mathscr{Z}_\varepsilon$, *a model* (y_1, X_1) *of minimal degree. If there exists more than one solution with the same degree, find one whose error*

$$\|y(t) - y_1(t)\| \tag{3.2}$$

is minimal.

PROBLEM 2. *Given a model* (y, X), *and* $k \geqq 0$, *find an* $\mathscr{Z} \cap \mathscr{Z}_k$, *a model* (y_2, X_2) *such that the quantity*

$$\|z(t) - y(t)\| \tag{3.3}$$

is minimized. If there exist more than one solution with the same error find one whose degree is minimal.

At first sight, it may seem reasonable to choose the class $\mathscr{Z}$ to be $\mathscr{Z}_\varepsilon$ (or $\mathscr{Z}_k$). This choice can certainly be made, but in this case we have to deal with a nonlinear minimization problem in L^2 over a set of rational functions (which is not even convex), and therefore this is quite a difficult problem to solve. We can then restrict the set of our candidates $\mathscr{Z}$ to some smaller and nicer class for which an explicit solution can be determined (as for the procedure proposed in this paper), or for which at least an approximate solution and an error bound can be computed (as for the Hankel-norm). In both cases we obtain an overall bound on the error of the actual solution. It is then clear that the performance of these algorithms will depend on the properties of the underlying set $\mathscr{Z}$.

DEFINITION. We say that an approximation algorithm is consistent if, whenever the smallest class $\mathscr{Z}_k$ containing the solution (y_1, X_1) also contains a minimal realization of y, then $y_1 = y$ almost everywhere, and (y, X_1) is also a minimal realization of y.

In other words, our algorithm should select a *minimal* exact model whenever there exists one of the same degree as in the solution.

In the Hankel-norm approximation scheme described in § 2, the set $\mathscr{Z}$ is, given (y, X) and k,

$$\mathscr{Z}_H := \{(z, Z) \in \mathscr{Z}_k ; Z^- = X^-\}.$$

The following example shows that this is not always a satisfactory choice.

Example 3.1. Let $X \cong (u, \bar{u})$ be any Markovian splitting subspace such that u is not equal to $\bar{u}$ and apply the Hankel-norm approximation to the model $(\bar{u}, X)$. Observe first that there exists a zero-dimensional exact realization of $\bar{u}$, namely $(\bar{u}, 0)$. Clearly we can obtain a zero-dimensional model using the Hankel norm, but this would not be exact. In fact, Hankel-norm approximation amounts to approximating the spectral factor of 1 (note that $\bar{u}$ is a white noise) corresponding to u, namely the structural function K/z (of degree, say, n) which is inner. The Hankel operator with symbol K/z has all its singular values equal to 1, and therefore the Hankel-norm approximant of K/z of degree $k < n$ will in fact be zero with error bound 1; then the corresponding model is $(0, 0)$, which clearly is not exact. Therefore we do not have consistency.

The example above shows that the choice of the original state space X is quite relevant to the quality of the Hankel-norm approximation. In fact, it is shown in [16] that the best behaviour is obtained by choosing the minimum-phase model. However, the given physical model may be far away from the minimum-phase one, and in this case a different approach is needed.

DEFINITION. We say that an algorithm has the inclusion property if for any solution (y_1, X_1) the relation $X_1 \subset X$ holds.

This property insures that in an appropriate basis the approximate model is a subsystem of the given one, as we will see below. In addition, as we shall show in § 5, *imposing* this property, namely choosing

$$\mathscr{Z}_I := \{(z, Z);\ Z \subset X\} \tag{3.4}$$

(where I stands for inclusion), will yield consistency. Therefore the Hankel-norm approximation cannot have the inclusion property, which is also seen from Example 9 in [7].

Desai and Pal [5] have suggested an approach to model reduction that has the inclusion and consistency properties and may, at first sight, seem very natural. However, as we shall see below, it yields a "model" that is not necessarily Markovian, and hence is not a model in our sense.

Let (y, X) be our given model. To simplify matters, we assume throughout the rest of this section that $X \subset H^-$ (the argument can be extended to any X with some slight modifications). Then X is splitting for y if

$$E^X \xi - E^{H^-} \xi = 0, \qquad \xi \in H^+, \tag{3.5}$$

i.e., the conditional angle between past and future given X is zero. It can be shown that $X_- = E^{H^+} H^-$ is the minimal subspace in H_- satisfying (3.5).

If a further reduction of X is needed, the splitting condition (3.5) must be waived. The natural way to do it is to impose that the angle between past and future, if not zero, is at least smaller than some given ε. That is, find $X_1 \subset H^-$ such that

$$\| E^{X_1} \xi - E^{H^-} \xi \| \leqq \varepsilon \| \xi \|, \qquad \xi \in H^+. \tag{3.6}$$

Mapping H^- onto $z^{-1}H^2$ and H^+ onto $\bar{H}^2$ via I_{u_-} and $I_{\bar{u}_+}$, respectively, the projection operator $E^{H^-} H^+$ has, as spectral representator, the Hankel operator H_T, where $T := z^{-1} W_- \bar{W}_+^{-1}$ is called phase function. Then the equivalent of (3.6) reads

$$\| E^{\mathscr{X}_1} M_T - E^{z^{-1}H^2} M_T \| \leqq \varepsilon, \tag{3.7}$$

$\mathscr{X}_1 \subset z^{-1}H^2$. It can be shown (see, e.g., [3], [5]) that if $\varepsilon = \sigma_k(H_T)$, and (ξ_i, η_i) is the Schmidt pair associated to $\sigma_i(H_T)$, then

$$\mathscr{X}_k := \operatorname{span}\{\xi_0, \cdots, \xi_{k-1}\}$$

satisfies condition (3.7). In fact, with a little computation, it can also be seen that

$$\langle \xi_k, \eta_k \rangle = \sigma_k(H_T)$$

and $\langle \xi_i, \eta_j \rangle = 0$ for $i \neq j$. In other words the value $\sigma_k(H_T)$ represents the cosine of the kth principal angle between H^+ and H^-, and is called the kth canonical correlation coefficient (see [2]). (The kth principal angle between two Hilbert spaces H, K is defined as

$$\arccos \inf \left\{ \sup_{\substack{\xi \in H_0, \eta \in K \\ \|\xi\| = \|\eta\| = 1}} < \xi, \eta;\ \operatorname{cod} H_0 = k,\ H_0 \subset H \right\}$$

and indicates roughly how near H and K are to being orthogonal, "given" $H \ominus H_0$.) Mapping $\mathscr{X}_1$ back to H^- via $I_{u_-}^{-1}$, we obtain an $X_1 \subset H^-$ satisfying (3.6). This representation is, in fact, the one obtained by Akaike using canonical correlation analysis [2]. It is rather straightforward to check that the canonical correlation coefficients are the

singular values of the Hankel operator H_T, and that the principal components of past and future are the inverse images under I_u of (ξ_k, η_k) (see [2]). The reason the procedure above looks interesting is that we can easily compute the approximate model from a realization of the original one, since it is always possible to find a basis in X such that the representation

$$x(t+1) = Ax(t) + Bu(t), \qquad y(t) = Cx(t) + Du(t)$$

of y induced by X has as its first k components exactly the images $I_{u-}^{-1}\eta_i$, for $i = 0, \cdots, k-1$, of the first k Schmidt pairs. In other words, we can write

$$\begin{bmatrix} x_1(t+1) \\ x_k(t+1) \\ x_{k+1}(t+1) \\ x_n(t+1) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_k(t) \\ x_{k+1}(t) \\ x_n(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_k \\ b_{k+1} \\ b_n \end{bmatrix} u(t),$$

$$y(t) = [c_1, \cdots, c_k \quad c_{k+1}, \cdots, c_n] \begin{bmatrix} x_1(t) \\ x_k(t) \\ x_{k+1}(t) \\ x_n(t) \end{bmatrix} \tag{3.8}$$

where $x_i(0) = I_{u-}^{-1}\eta_i$. Then an approximate model is naturally given by

$$\begin{bmatrix} x_1(t+1) \\ x_k(t+1) \end{bmatrix} = A_{11} \begin{bmatrix} x_1(t) \\ x_k(t) \end{bmatrix} + \begin{bmatrix} b_1 \\ b_k \end{bmatrix} u(t),$$

$$y(t) = [c_1, \cdots, c_k] \begin{bmatrix} x_1(t) \\ x_k(t) \end{bmatrix}. \tag{3.9}$$

This is the approach already suggested by Akaike in [2], and proposed by Desai and Pal [5]. The main drawback, however, is that (3.9) is indeed a submodel of (3.8) only if A_{12} is zero, and this generally cannot be achieved if we require the first k components to be the images of the η_i, $i = 0, \cdots, k-1$.

An equivalent statement is that X_1 is, in general, not Markovian. In fact, we are going to prove that X_1 cannot be Markovian if the singular values of H_T are distinct.

THEOREM 3.1. *Let* $(\xi_0, \eta_0), \cdots, (\xi_{n-1}, \eta_{n-1})$ *be the Schmidt pairs of* H_T *associated to* $\sigma_0 > \sigma_1 > \cdots > \sigma_{n-1}$. *Then* $X_k = T_u^{-1}$ span $\{\eta_0, \cdots, \eta_{k-1}\}$ *is not Markovian, for* $k < n-1$.

The proof is quite technical and is given in the Appendix.

The basic reason this algorithm does not yield a Markovian model is that what is reduced is simultaneously almost unobservable *and* almost unconstructible (this is the meaning of (3.6)). By contrast, the reduction algorithm of Theorem 2.1 cuts off the unobservable part of X. That is, letting $\bar{S}_1 := S^{\perp} \vee H^+$ and $\bar{S} \ominus \bar{S}_1 = X_u$ easily shows that $E^{H^+}X_u = 0$. But X_u is certainly not unconstructible, because $X_u \subset (H_1)^{\perp}$ implies $E^{X_u}H_- = X_u$, since $H^- \vee H^+ = H$. So in fact X_u is constructible.

We conclude that the philosophy of the solution to the model approximation problem is to take observability and constructibility into account separately, not at the same time. We will do this in the next two sections. It is easily seen that a procedure for solving Problem 1 can be easily transformed into a solution to Problem 2. Hence, in what follows, we will focus our attention on Problem 1.

4. The exact model. As we have said, we want a model reduction scheme that is consistent. To this end, we reformulate the exact reduction algorithm of Theorem 2.1

so that it can be easily generalized for approximation. This is the content of this section and is, in particular, the meaning of Theorem 4.2 and Algorithm 4, as Example 4.1 shows.

There are basically three variations of the algorithm of Theorem 2.1 when $X \cong (S, \bar{S})$ is given.

ALGORITHM 1. Set $\bar{S}_1 := S^\perp \vee H^+$, $S_1 := \bar{S}_1^\perp \vee H^+$.

We obtain the minimal state space contained in X that is closest to the future H^+ in the sense that X_1 is the maximal element in the complete sublattice of minimal splitting subspaces of X. In particular, if the backward predictor space $X_+ = E^{H^+}H^-$ is contained in X, then $X_1 = X_+$.

ALGORITHM 2. Set $S_1 := \bar{S}^\perp \vee H^+$, $S_1 := S^\perp \vee H^+$.

We get the subspace that is closest to the past in the above ordering. In particular, if $X \supset X_- = E^{H^-}H^+$, the predictor space, then $X_1 = X$.

ALGORITHM 3. Take S_1 to be any invariant subspace with the property

$$H^- \vee \bar{S}^\perp \subset S_1 \subset S_+ \cap S$$

and set $\bar{S}_1 := S_1^\perp \vee H^+$. Then clearly S_1 and $\bar{S}$ intersect perpendicularly and it can be shown that X_1 is minimal.

In the following we will focus our attention on the first two algorithms.

The frame space $X^\square$ is defined as

$$X^\square := X_- \vee X_+$$

where X_- and X_+ are the forward and backward predictor spaces. It can be shown that any minimal splitting subspace is contained in $X^\square$ [11]. Moreover, these subspaces have a complete lattice structure under the partial ordering described below. It can be shown [11] that if X is minimal splitting,

$$X = \int Q^* H(K)\, d\hat{u}_-$$

where Q is an inner function. The correspondence between X and Q is one to one.

Then given X_1 and X_2 minimal splitting and the corresponding inner function Q_1, Q_2, we say that $X_1 < X_2$ if $Q_1 | Q_2$.

THEOREM 4.1. *Suppose X Markovian splitting subspace for y and set $X_1 := X \cap X^\square$. Then the set of minimal subspaces of X_1 is a complete lattice with the ordering induced by $X^\square$ (i.e., it has a complete lattice structure). The maximal element is given by Algorithm* 1 *and the minimal element by Algorithm* 2.

For the proof we need the following lemma.

LEMMA 4.1. *Suppose $Z \subset X^\square$ is a Markovian splitting space with a complete lattice structure. Then $R \cap Z$ also has a complete lattice structure for any left invariant subspace R.*

Proof. Let $Z_- \cong (v_-, \bar{v}_-)$ be the minimal element in the lattice and Z_+ the maximal. Then [11]

$$Z_- = \int H(K)\, d\hat{v}_-, \qquad Z_+ = \int Q_+^* H(K)\, d\hat{v}_-$$

and any minimal subspace of Z has the form

$$X = \int Q^* H(K)\, d\hat{v}_-$$

for some $Q \mid Q_+^*$. Conversely, to any such Q there corresponds a minimal state space. The space R will have the representation

$$R = \int P^* H^2 \, d\hat{v}_-$$

for some P inner. Then set

$$Q_m := (Q_+, P).$$

Then

$$X_m := \int Q_m^* H(K) \, d\hat{v}_-$$

is the maximal element of our sublattice, since the inner divisors of an inner function form a complete sublattice. The minimal element is clearly Z_-, and the lemma is proven. □

Proof of Theorem 4.1. Since $X \cong (S, \bar{S})$ we can apply the lemma above to $X^\square \cap S$, and then (after time reversal) to $(X^\square \cap S) \cap \bar{S}$, and the first part of the theorem is thus proven.

To see that the minimal element is the one given by Algorithm 2, observe that $X_1 < X_2$ in the lattice if and only if $Q_1 \mid Q_2$, where

$$X_i = \int Q_i^* H(K) \, du_-,$$

and hence if $S_1 \subset S_2$, since

$$S_i = \int Q_i^* H^2 \, du_-.$$

Since $S_i \supset \bar{S}^\perp \vee H^-$, clearly the minimal element is the one for which $S = \bar{S}^\perp V H^-$. This is similar for the maximal element. □

We want to describe the reduction procedure of Theorem 2.1 in the spectral domain. Therefore we introduce a basis in the spectral image $\mathscr{X}$ of the given state space X, and hence also in X. Let I_u denote the isomorphism mapping X^- onto $z^{-1}H^2$. Then (cf. (2.13), (2.14))

$$X = \int z^{-1} H(K) \, d\hat{u}$$

where K is inner. Let $b_1, \cdots, b_n$ denote the poles of K. Then a basis for $\mathscr{X} = z^{-1}H(K)$ is

$$\begin{aligned} v_1(z) &= \frac{N_1}{z - b_1}, \\ v_2(z) &= \frac{N_2}{z - b_2} \frac{1 - z\bar{b}_1}{z - b_1}, \\ v_n(z) &= \frac{N_n}{z - b_n} \frac{1 - z\bar{b}_{n-1}}{z - b_{n-1}} \cdots \frac{1 - z\bar{b}_1}{z - b_1} \end{aligned} \tag{4.1}$$

where N_i is a normalizing factor given by $N_i = (1 - |b_i|^2)^{1/2}$. The basis $v_1, \cdots, v_n$ is orthonormal. For short we can write

$$v_i(z) = \frac{N_i}{z - b_i} \prod_{j=1}^{i-1} B_j(z), \qquad B_j(z) := \frac{1 - z\bar{b}_j}{z - b_j}.$$

The basis $v_1, \cdots, v_n$ also has the property that $\operatorname{span}\{v_1, \cdots, v_n\} = z^{-1}H(B_1, \cdots, B_i)$. This basis clearly depends on the ordering of the poles. If σ is a permutation of S_n, then $b_{\sigma_1}, \cdots, b_{\sigma_n}$ will generate another basis, i.e.,

$$v_i^\sigma(z) := \frac{N_{\sigma_i}}{z - b_{\sigma_i}} \prod_{j=1}^{i-1} B_{\sigma_j}(z).$$

The spectral factor W and the isomorphic image of $y(0)$ under I_u then have different representations as σ ranges over S_n:

$$W = \sum_{i=0}^{n} \alpha_i(\sigma) v_i^\sigma \tag{4.2}$$

where $v_0^\sigma = 1$. Let $\operatorname{Res}(v_i, b_j)$ denote the residue of the function $v_i(z)$ in b_j. Then the coefficients $\alpha_i(\sigma)$ are the solution to the following system:

(4.3)

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \operatorname{Res}(v_1^\sigma; b_{\sigma_1}) & \operatorname{Res}(v_1^\sigma; b_{\sigma_1}) & \cdots & \operatorname{Res}(v_n^\sigma; b_{\sigma_1}) \\ 0 & 0 & \operatorname{Res}(v_2^\sigma; b_{\sigma_2}) & \cdots & \operatorname{Res}(v_n^\sigma; b_{\sigma_2}) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \operatorname{Res}(v_n^\sigma; b_{\sigma_n}) \end{bmatrix} \begin{bmatrix} \alpha_0(\sigma) \\ \alpha_1(\sigma) \\ \alpha_2(\sigma) \\ \vdots \\ \alpha_n(\sigma) \end{bmatrix} = \begin{bmatrix} W(\infty) \\ \operatorname{Res}(W; b_{\sigma_1}) \\ \operatorname{Res}(W; b_{\sigma_2}) \\ \vdots \\ \operatorname{Res}(W; b_{\sigma_n}) \end{bmatrix}$$

and can be computed recursively (cf. [3]).

Among all possible choices of basis, there are some that are particularly interesting.

LEMMA 4.2. *Let $X \cong (W, \bar{W}) \cong (u, \bar{u})$. Then X is observable if and only if there exists no permutation $\sigma \in S_n$ subject to $\alpha_n(0) = 0$ in the representation* (4.2) *of W.*

Proof (If). Suppose that, for some σ, $\operatorname{Res}(W; b_{\sigma_n}) = 0$. This means $\alpha_n(\sigma) = 0$, and hence the degree of W is at most $n-1$. It can be shown (cf. [11]) that, given $X \cong (W, \bar{W})$, there exist polynomials p and q such that $W(z) = p(z)|q(z)$ and $\bar{W}(z) = p(z)|\hat{q}(z)$, where $\hat{q}(z) = z^n q(z^{-n})$, $n = \deg q$. Moreover, $X \cong (W, \bar{W})$ is observable if p and q are coprime (see [11]). In our case, $\deg q = n$ and $\deg W \leqq n-1$ imply that p and q are not coprime, i.e., X is not observable. Reading the argument backward proves the (Only if).

In other words Lemma 4.2 characterizes the observable subspaces of X in the spectral domain. Clearly an analogous result holds for constructibility. To obtain it we simply need to consider the map $I_{\bar{u}}$ from X^+ onto $\bar{H}^2$. Then the conjugate basis of v_i^σ for $i = 1, \cdots, n+1$, is

$$\bar{v}_i^\sigma(z) := \frac{zN_{\sigma_i}}{1 - zb_{\sigma_i}} \prod_{j=i+1}^{n+1} B_{\sigma_j}^*(z), \qquad B_j(z) := \frac{1 - z\bar{b}_j}{z - b_j}. \tag{4.4}$$

(The $(n+1)$st component represents the constants. This is done to simplify notation later.) The *conjugate* spectral factor of y will therefore have the representation

$$\bar{W} = \sum_{i=1}^{n+1} \bar{\alpha}_i(\sigma) \bar{v}_i^\sigma. \tag{4.5}$$

Then the constructible version of Lemma 4.2 reads as follows.

LEMMA 4.3. *Let $X \cong (W, \bar{W}) \cong (u, \bar{u})$. Then X is constructible if and only if there exists no permutation $\sigma \in S_n$ subject to $\bar{\alpha}_1(\sigma) = 0$ in the representation* (4.5) *of $\bar{W}$.*

The reason Lemmas 4.2 and 4.3 are interesting is that they provide the key to the spectral version of Algorithm 2 (the reduction algorithm) as is shown in the next theorem.

First define, for a given Markovian subspace X splitting for y,

$$k_o := \min_{\sigma \in S_n} \{k;\ \alpha_j(\sigma) = 0, j > k\}, \tag{4.6}$$

$$k_c := \max_{\sigma \in S_n} \{k;\ \bar{\alpha}_j(\sigma) = 0, j < k\}. \tag{4.7}$$

Let σ_o and σ_c denote the minimizer and maximizer, respectively, in (4.6) and (4.7). The letters o and c here stand for observable and constructible, respectively.

LEMMA 4.4. *Let $X \cong (u, \bar{u})$. Then*

$$X_o := I_u^{-1} \operatorname{span} \{v_1^{\sigma_o}, \cdots, v_{k_o}^{\sigma_o}\} \tag{4.8}$$

and

$$X_c := I_{\bar{u}}^{-1} z^{-1} \operatorname{span} \{\bar{v}_{k_c}^{\sigma_c}, \cdots, \bar{v}_n^{\sigma_c}\} \tag{4.9}$$

are, respectively, observable and constructible Markovian splitting subspaces for y.

Before proving this lemma, we proceed to the main theorem of this section.

THEOREM 4.2. *Let $X \cong (S, \bar{S})$ be Markovian splitting for y, let X_o and X_c be defined as in Lemma* 4.4, *and set $\bar{S}_1 := S^{\perp} \vee H^{+}$. Then*

$$(S, \bar{S}_1) \cong X_o.$$

Analogously, let $S_1 := \bar{S}^{\perp} \vee H^{-}$. Then

$$(S_1, \bar{S}) \cong X_c.$$

The spectral version of Algorithm 1 is now clear.

ALGORITHM 4. Let $X \cong (u, \bar{u})$ be given.

(1) Set $X_o := I_u^{-1} \operatorname{span} \{v_1^{\sigma_o}, \cdots, v_{k_o}^{\sigma_o}\}$.

(2) Consider $X_o \cong (W, \bar{W}_o) \cong (u, \bar{u}_o)$, the basis $I_{\bar{u}}^{-1} z^{-1} \{\bar{v}_i^{\bar{\sigma}_o}\}$ in X_o, where $\bar{\sigma}_o \in S_{k_o}$, and the representation

$$\bar{W}_o = \sum_{i=1}^{k_o+1} \bar{\alpha}_1(\bar{\sigma}) \bar{v}_i^{\bar{\sigma}_o}.$$

(3) Set $k_{oc} := \max_{\sigma_o \in S_{k_o}} \{k\colon \bar{\alpha}_j(\bar{\sigma}_o) = 0, j < k\}$ and let σ_{oc} be the maximizer of this expression.

(4) Define $X_{oc} := I_{\bar{u}_o}^{-1} z^{-1} \operatorname{span} \{v_i^{\sigma_{oc}};\ i = k_{oc}, \cdots, k_o\}$.

Obviously, there is an analogous spectral domain version of Algorithm 2, yielding a minimal Markovian splitting subspace X_{co}.

COROLLARY 4.1. *Let $S_1, \bar{S}_1$, as in Algorithm* 1. *Then*

$$X_{oc} \cong (S_1, \bar{S}_1). \tag{4.10}$$

Analogously, if $S_2, \bar{S}_2$ are as in Algorithm 2, *then*

$$X_{co} \cong (S_2, \bar{S}_2).$$

With this corollary the spectral equivalents of Algorithms 1 and 2 are completely characterized. Before we look at the proofs, we consider an example, trivial by itself but useful for the understanding of the approximation procedure in the next chapter.

Example 4.1. Consider $X \cong (u, \bar{u})$:

$$X = I_u^{-1} \operatorname{span}\left\{\frac{1}{z+(1/2)}, \frac{1}{z+(4/5)}\right\} = I_{\bar{u}}^{-1} \operatorname{span}\left\{\frac{1}{1+(z/2)}, \frac{1}{1+z(4/5)}\right\}$$

and

$$y(0) = \int W(e^{i\omega})\, du(\omega)$$

with

$$W(z) = \frac{z+(1/3)}{1+(z/2)} \frac{1+z(4/5)}{z+(4/5)}.$$

Then $\bar{W}(z)$ can be written as

$$\bar{W}(z) = \frac{z+(1/3)}{1+(z/2)} \frac{1+z(4/5)}{1+z(4/5)}.$$

It is easily seen that, whereas no reduction is possible on W, for $\bar{W}$ there are two representations associated to two different bases:

$$\bar{W}(z) = \frac{1}{3} + \frac{1}{2} \frac{z}{1+z(4/5)} + \frac{5}{12} \frac{z}{1+(z/2)} \frac{z+(4/5)}{1+z(4/5)},$$

$$\bar{W}(z) = \frac{1}{3} + \frac{5}{6} \frac{z}{1+(z/2)} + 0 \frac{z}{1+z(4/5)} \frac{z}{1+(z/2)}.$$

The second representation shows that, according to Lemma 4.3, X is not constructible, because α_2 is zero, and hence it can be reduced in the second step of the algorithm. Then $X_0 = X$ and no reduction is made, whereas

$$X_{oc} = I_{\bar{u}}^{-1}\left\{\frac{1}{1+(z/2)}\right\}$$

is a one-dimensional minimal splitting subspace for y.

As we said, it is obvious that $\bar{W}$ can be represented by a rational function of first degree. It is important to construct the reduced state space in a way that is generalizable to approximation, as will be seen in the next chapter.

We turn now to the proofs. We consider only the observable case, the constructible one being symmetric.

LEMMA 4.5. *Let $X \cong (u, \bar{u})$ be a Markovian subspace. If for some constant d*

$$y(0) = E^X y(0) + du(0) \tag{4.11}$$

then X is splitting for y. Similarly, if $\mathscr{X}$ and W are the spectral representations of X and $y(0)$, X is splitting for y if for some constant d

$$W \in d + \chi.$$

Proof. Let $E^X y(t) = Cx(t)$, where C is a convenient matrix and $x(t) = U^t x(0)$, $x(0) \in X$. Since for $t \geqq 0$, $u(t) \perp X^-$, we have $E^{X^-} y(t) = E^{X^-} Cx(t)$ because X is Markovian. Similarly, for $t < 0$, since $u(t) \perp (X^-)^\perp = X^+ \ominus X$ from Theorem 2.3, we can write

$$E^{X^+} y(t) = E^{X^+} Cx(t) + E^{X^+} u(t) = E^X Cx(t) + E^X u(t) = E^X y(t).$$

In other words, $E^{H^- \vee X} H^+ = E^X H^+$ and $E^{H^+ \vee X} H^- = E^X H^-$, i.e., X is splitting. The second part follows by mapping (4.11) in the spectral domain.

Proof of Lemma 4.4. Let $X_o \cong (u_o, \bar{u}_o)$ and consider in H^2 the images $\mathscr{X}_o$ and W of X_o and $y(0)$ under the map I_u. Then Markovianness follows from the left invariance of $\mathscr{X}_o$. As for the splitting property, from (4.2) we see that $W \in d + \mathscr{X}$, and we can therefore apply Lemma 4.5. As for observability note first that, since X_o is splitting, the representation $X_o \cong (W_o, \bar{W}_o)$ holds for a convenient choice of spectral factors W_o, $\bar{W}_o$. We need to prove that $W_o = W$, or equivalently, that $S = X_o^-$. This equality becomes, in the spectral domain, under the map I_u,

$$\operatorname{span}\{z^n \mathscr{X}_o;\ n \leqq 0\} = z^{-1}H^2.$$

To prove this last equality we pick a cyclic element in $\mathscr{X}_o$. Consider the first element of the basis of $\mathscr{X}_o$, $v_1 = N_1(z - b_1)^{-1} = N_1 z^{-1}(1 - z^{-1}b_1)^{-1}$. The function $(1 - z^{-1}b_1)^{-1}$ is invertible in H^2, and hence outer, i.e., $v_1 H^2 = z^{-1}H^2$, as wanted. That is, $S = X_o^-$, and hence $W_o = W$. But then we can apply Lemma 4.2, and X_o is observable.

Proof of Theorem 4.2. Let $X_o \cong (S, \bar{S}_o)$, and $\bar{S}_1 := H^+ \vee S^\perp$. We need to show only that $\bar{S}_o = \bar{S}_1$. Since X_o is splitting, we have $H^+ \subset \bar{S}_o$, and since also $\bar{S}_o \supset S^\perp$, we obtain $\bar{S}_o \supset \bar{S}_1$. That is, $X_o \supset X_1 = (S, \bar{S}_1)$. Since both X_o and X_1 are observable subspaces, the following holds:

$$E^S H^+ = E^{X_1} H^+ = X_1 \subset X = E^X H^+ = E^S H^+, \tag{4.12}$$

i.e., $X_1 = X$. A similar argument holds for X_c.

5. Main results. We have seen that, if X is not minimal, there exists an algorithm to cut out the unobservable and unconstructible parts of X. Suppose now that we are in a situation where a minimal model is still too large for applications. Then some further reduction leading to an approximate model must be performed. This is the topic of this section.

The idea is to generalize the algorithm presented above by cutting off subspaces of X that are "almost" unobservable or "almost" unconstructible, in a specific sense to be described below.

To this end, let us examine Algorithm 1 more closely.

LEMMA 5.1. *Let* $X \cong (S, \bar{S})$ *and* $\bar{S}_o := S^\perp \vee H^+$. *Then* $X_o \cong (S, \bar{S}_o)$ *is characterized by the following properties*:

(5.1)
- (a) X_o *is Markovian*;
- (b) $X \ominus X_o$ *is unobservable*;
- (c) $X^- = X_o^-$;
- (d) X_o *is minimal with respect to the above properties* (*i.e., if* $X'_o \subset X_o$ *has properties* (a)-(c), *then* $X'_o = X_o$).

Proof. X_o is Markovian by construction, and since $X^- = S = X_1^-$, then (c) is also satisfied. Since $(X \ominus X_o) \perp (\bar{S}_o \cap S)$, but also $(X \ominus X_o) \subset S$, it follows that $(X \ominus X_o) \perp \bar{S}_o$. Since $\bar{S}_o \supset H^+$, a fortiori $(X \ominus X_o) \perp H^+$, i.e., $X \ominus X_o$ is unobservable, and this shows (b). Suppose now that $X'_o \subset X_o$ has the same properties. Then $(X \ominus X'_o) \cap X_o$ is also unobservable. But this implies $(X \ominus X'_o) \cap X_o = 0$, i.e., $X'_o = X_o$ and hence (d) holds.

Conversely, if a subspace $X_1 \subset X$ satisfies (a), (b), (c), (d), then it is splitting, because for each $\xi \in H^+$, the equality

$$E^X \xi = E^{X_1} \xi + E^{X \ominus X_1} \xi = E^{X_1} \xi$$

holds and, since X is splitting, we have

$$E^S \xi = E^X \xi = E^{X_1},$$

which is the splitting property for X_1. Observability follows from (d). In conclusion, X_1 is Markovian, observable, and splitting for y, and such that $X_1 = S$. But then, $X_1 = X_o$ (because $X_1^- = S$ and $X_1^+ \supset H^+ \vee S^\perp$). Using (4.12), we get $X_1 = X_o$.

For our approximation scheme, we need to replace the unobservability condition (b) by a weaker one (which will not imply, as condition (b) did, that X_o is splitting).

DEFINITION. Let $\varepsilon > 0$ be given. Then we say that a subspace Z of H is ε-unobservable if

$$\sup \|E^Z y(t)\| \leqq \varepsilon, \qquad t \geqq 0 \tag{5.2}$$

and ε-unconstructible if

$$\sup \|E^Z y(t)\| \leqq \varepsilon, \qquad t < 0. \tag{5.3}$$

The reason for this definition is explained by the following proposition.

PROPOSITION 5.1. *$Z \subset X$ is unobservable if it is ε-unobservable for all $\varepsilon > 0$.*

Proof. Z is unobservable if $Z \subset X \cap (H^+)^\perp$. But this implies that $E^Z x = 0$ for $x \in H^+$. In particular, $E^Z y(t) = 0$ for all t. Conversely, ε-unobservability of Z for all $\varepsilon > 0$ implies $E^Z y(t) = 0$ for $t > 0$, and since any $x \in H^+$ has a representation $x = \sum_{t \geqq 0} \alpha_t y(t)$, $E^Z y(t) = 0$ for $t \geqq 0$, we get $E^Z x = 0$. □

Condition (b) now becomes the following:

(b′) $X \ominus X_1$ is ε-unobservable.

THEOREM 5.1. *Suppose $X_1 \subset X$ satisfies condition* (a)-(c), *and $X \cong (u, \bar{u})$. Then there exists a process $\{y_1(t)\}$ such that*

$$\|y(t) - y_1(t)\| < \varepsilon \tag{5.4}$$

and X_1 is splitting for y_1. If condition (d) *is satisfied then X_1 is observable for y_1.*

Proof. Since X is splitting for y there exists a representation

$$y(0) = du(0) + E^X y(0)$$

where u is the forward process associated to X. Set

$$y_1(0) := du(0) + E^{X_1} y(0).$$

Then,

$$y_1(0) = du(0) + E^{X_1} y_1(0).$$

So, from Lemma 4.5, X_1 is splitting for y_1. Moreover, $X \ominus X_1$ is ε-unobservable, and hence

$$\begin{aligned} \|y(0) - y_1(0)\| &= \|E^X y(0) - E^{X_1} y(0)\| \\ &= \|E^{X \ominus X_1} y(0)\| < \varepsilon. \end{aligned}$$

Let us now assume that X_1 is not observable with respect to y_1. We shall show that condition (d) cannot then hold. If $X_1 \cong (S, \bar{S}_1)$ is not observable, then $X_2 \cong (S, S^\perp \vee H^+(y_1))$ is a proper subspace of X_1 satisfying properties (a), (b′), and (c); hence (d) cannot hold. □

Now we exhibit a constructive way to obtain a reduced space satisfying properties (a), (b′), (c), and (d). First the following lemma is needed.

LEMMA 5.2. *Let $X_1 \cong (S_1, \bar{S}_1) \subset X \cong (S, \bar{S})$. Then $\bar{S} \ominus \bar{S}_1$ is ε-unobservable if and only if*

$$\|E^{\bar{S} \ominus \bar{S}_1} y(0)\| < \varepsilon. \tag{5.5a}$$

Similarly, $S \ominus S_1$ is ε-unconstructible if and only if

$$\|E^{S \ominus S_1} y(0)\| < \varepsilon. \tag{5.5b}$$

Proof. Suppose (5.5a) holds, and let $y(0) = y_1(0) + y_2(0)$ with $y_1(0) \in \bar{S}_1$ and $y_2(0) \in \bar{S} \ominus \bar{S}_1$. Then $E^{\bar{S}\ominus\bar{S}_1} y_1(t) = 0$ and

$$\begin{aligned} \|E^{\bar{S}\ominus\bar{S}_1} y(t)\| &= \|E^{\bar{S}\ominus\bar{S}_1} y_2(t)\| \\ &= \|E^{\bar{S}\ominus\bar{S}_1} U' y_2(0)\| \leqq \|E^{\bar{S}\ominus\bar{S}_1} y_2(0)\| = \|y_2(0)\| \\ &= \|E^{\bar{S}\ominus\bar{S}_1} y(0)\| \end{aligned} \qquad t \geqq 0$$

because U is unitary and the projection is contractive. Hence $\bar{S} \ominus \bar{S}_1$ is ε-unobservable.

The other direction is obvious. A similar argument holds for constructibility.

THEOREM 5.2. *Let $X \cong (u, \bar{u})$ be given, and let $\mathscr{X} := I_u X$. Suppose there exists a $\sigma \in S_n$ such that $\sum_{i=k+1}^{n} \|\alpha_i(\sigma)\|^2 \leqq \varepsilon^2$. Then $X_\sigma = \operatorname{span}\{v_1^\sigma, \cdots, v_k^\sigma\}$ satisfies conditions* (a), (b′), *and* (c). *If $\sum_{i=k}^{n} |\alpha_i(\sigma)|^2 > \varepsilon^2$ for each $\sigma \in S_n$, then condition* (d) *is also satisfied.*

Proof. Clearly X_σ is Markovian. The space X^- is mapped under I_U onto $z^{-1}H^2$. Since $v_1^\sigma = z^{-1}(1 - zb_{\sigma^1})^{-1}$ spans $z^{-1}H^2$ (i.e., $\overline{\operatorname{span}}\{z^{-n}v_1;\ n \geqq 0\} = z^{-1}H^2$); also $X_\sigma^- = X^-$. To prove that condition (b′) holds, we apply Lemma 5.2 to X and X_σ. Concerning condition (d), if $X_{\sigma^1} \subset X_\sigma$, for some $\sigma^1 \in S_{k^1}$, with $k^1 < k$, and X_{σ^1} also satisfies conditions (a), (b′), and (c), then by definition $\sum_{i=k^1+1}^{n} |\alpha_i(\sigma^1)|^2 \leqq \varepsilon^2$. But this contradicts the hypothesis of the theorem.

The following algorithm concludes this story.

ALGORITHM 5a. Let the state space $X \cong (u, \bar{u})$ and $\varepsilon \geqq 0$ be given. We set

$$k_o := \min\left\{k;\ \sum_{i=k+1}^{n} |\alpha_i(\sigma)|^2 \leqq \varepsilon^2 \text{ for some } \sigma \in S_n\right\} \tag{5.6}$$

and let σ_o be a minimizer of (5.6). (This minimizer does not need to be unique. In this case we choose the one for which the above sum in parentheses is minimal.) We set

$$X_o := I_u^{-1} \operatorname{span}\{v_1^{\sigma_o}, \cdots, v_{k_o}^{\sigma_o}\},$$

$$y_o(0) := \int \sum_{i=0}^{k_o} \alpha_i(\sigma) v_i^\sigma \, d\hat{u}.$$

Then, by Theorem 5.1, X_o is Markovian, splitting, and observable for y_o. Moreover,

$$\|y(0) - y_o(0)\| \leqq \varepsilon.$$

To complete the analogy with Algorithm 1, we need to cut out an ε-unconstructible part from the resulting space. This can be easily done by reversing time and using the conjugate objects $\bar{u}$, $\bar{H}^2$, etc. instead of u, H^2. In the following example, we consider the same state space as in Example 4.1 but use a slightly different process.

Example 5.1. Let $X \cong (u, \bar{u})$ be given by

$$X = I_u^{-1} \operatorname{span}\left\{\frac{1}{z + (1/2)}, \frac{1}{z + (4/5)}\right\},$$

$$y(0) = \int_{-\pi}^{\pi} W(e^{i\omega}) \, du(\omega)$$

where

$$W(z) = \frac{z + (1/3)}{z + (1/2)} \frac{z + (11/9)}{z + (4/5)}.$$

The corresponding $\bar{W}$ is

$$\bar{W}(z)=\frac{z+(1/3)}{1+(z/2)}\,\frac{z+(11/9)}{1+(4z/5)}.$$

It is simple to see that X is now minimal. Therefore a one-dimensional model needs to be an approximation with some error $\varepsilon>0$. Let $\varepsilon=0.05$. Then it is easily seen that X has no ε-unobservable part, and hence $X_0=X$. On the other hand, $\bar{W}$ can be written as

$$\begin{aligned}\bar{W}(z)&=\alpha_0+\alpha_1\bar{v}_1+\alpha_2\bar{v}_2\\&=\frac{11}{27}+\frac{164}{81\sqrt{3}}\,\frac{z(\sqrt{3}/2)}{1+(z/2)}+\frac{11}{243}\,\frac{5/3}{1+(4z/5)}\,\frac{z+(1/2)}{1+(z/2)}\end{aligned}$$

with $\bar{v}_1\perp\bar{v}_2$ and $\|\bar{v}_1\|_{L^2}=\|\bar{v}_2\|_{L^2}=1$. Since

$$\frac{11}{243}=.045267489$$

we obtain that span $I_u^{-1}\bar{v}_2$ is ε-unconstructible and can be eliminated. Therefore

$$X_{oc}=\text{span } I_{\bar{u}}^{-1}\left\{\frac{1}{1+z/2}\right\},$$

$$y_{oc}(0)=\int_{-\pi}^{\pi}\left(\frac{11}{27}+\frac{164}{81\sqrt{3}}\,\frac{z\sqrt{3/2}}{1+z/2}\right)d\hat{\bar{u}}.$$

Now we compare our approximation with one obtained using Hankel-norm techniques. The Hankel-norm approximation applied to the spectral factor W yields the approximant

$$A(z)=-.388370\,\frac{z-.986176}{z+.857631}$$

with error bound

$$\sigma_1=.167481,$$

which is considerably larger than our bound. An objection might be that we did not apply Hankel-norm approximation to the most favourable case, since W is not minimum-phase (cf. [16]). The first answer to this objection is that we are interested in an approximation of the original model, which includes the state space, and not in that of the corresponding minimum-phase model.

Anyway, for sake of completeness, we approximate the outer factor of W:

$$W_1(z)=\frac{11}{9}\,\frac{z+1/3}{z+1/2}\,\frac{z+9/11}{z+4/5}.$$

The Hankel-norm approximation of W_1 yields an error of

$$\sigma_2=\frac{40}{27}\,\frac{7}{108}\cong.096022,$$

which is still more than twice our bound. Moreover, again we stress that this is *not* an approximant of the original spectral factor.

However, Hankel-norm approximation provides, in general, an upper bound on the L^2 norm of the error, which might a priori be smaller. This is not the case in the above setting, since in both cases the difference between the function and the Hankel-norm approximant is inner (cf. [7]), and therefore the errors in L^2-norm, Hankel-norm, and H^∞-norm do coincide. Therefore the errors σ_1 and σ_2 are also real L^2-errors.

There remain some questions to be answered relating to the minimality of the final space. A first question is what the basis of the backward representation will look like. The backward representation $I_{\hat{u}}$ maps $\bar{S}$ onto $\bar{H}^2$, and

$$y(0) = \int \bar{W}\, d\hat{u}.$$

Now we want to express W as the sum of a constant and something in $z\bar{\mathscr{X}} = I_{\hat{u}} UX$ (see 2.11). The canonical basis for $z\bar{\mathscr{X}}$ is thus given by

$$\bar{v}_i^\sigma := \frac{N_{\sigma_i}}{z^{-1} - b_{\sigma_i}} \prod_{j=i+1}^{n} \frac{z - b_{\sigma_j}}{1 - z b_{\sigma_j}}$$

and the conjugate spectral factor $\bar{W}$ will have the representation

$$\bar{W} = \sum_{i=1}^{n+1} \beta_i(\sigma) \bar{v}_i^\sigma \tag{5.7}$$

where $\bar{v}_{n+1}^\sigma = 1$. The forward and the backward basis are related via the following lemma.

LEMMA 5.3. *Let $\alpha_i(\sigma)$ and $\beta_i(\sigma)$ be the coefficients of W and $\bar{W}$ in* (4.2) *and* (5.7), *respectively, related to the same permutation of the poles σ. Then*

$$\beta_i(\sigma) = \alpha_i(\sigma) b_{\sigma_i} + N_{\sigma_i} \left(\sum_{k=0}^{i-1} \alpha_k(\sigma) N_{\sigma_k} p_{k,i}(\sigma) \right), \qquad i = 1, \cdots, n,$$

$$\beta_{n+1}(\sigma) = \sum_{k=0}^{n} \alpha_k(\sigma) N_{\sigma_k} p_{k,n}(\sigma)$$

where

$$p_{k,i}(\sigma) = \prod_{j=k+1}^{i-1} (-b_{\sigma_j}) \qquad (p_{k,i} = 1 \text{ for } 1 \leqq k+1)$$

and $N_{\sigma_0} = 1$.

Proof. We drop the dependence on σ in the proof, understanding that what holds for $1, \cdots, n$, is also true for $\sigma_1, \cdots, \sigma_n$. Let

$$C_i := \prod_{j=i+1}^{n} B_j^*.$$

Then, remembering that $\bar{W} = K^* W$ (and $K^* = C_0$), we obtain

$$\begin{aligned} v_i K^* &= \frac{N_i}{z - b_i} (B_{i-1} \cdots, B_1) K^* \\ &= \frac{N_i}{1 - z b_i} (B_{i+1}^* \cdots B_n^*) \\ &= \frac{z N_i b_i}{1 - z b_i} (B_{i+1}^* \cdots B_n^*) + N_i (B_{i+1}^* \cdots B_n^*) \\ &= b_i \bar{v}_i + N_i C_i. \end{aligned} \tag{5.8}$$

Next, $C_{i-1} = N_i\bar{v}_i - b_iC_i$. In fact,

$$C_{i-1} = \frac{z-b_i}{1-zb_i}\,C_i = \frac{1-z^{-1}b_i}{z^{-1}-b_i}\,C_i$$
$$= \left(\frac{1-\bar{b}_ib_i}{z^{-1}-b_i} - b_i\right)C_i = N_i\bar{v}_i - b_iC_i,$$

and therefore the following representation holds:

$$C_k = \sum_{i=k+1}^{n} N_i\bar{v}_i\left(\prod_{j=k+1}^{i-1}(-b_j)\right) + \prod_{j=k+1}^{n}(-b_j) \tag{5.9}$$
$$= \sum_{i=k+1}^{n} N_i\bar{v}_ip_{k,i} + p_{k,n+1}.$$

Hence, expressing $\bar{W}$ through (5.8) and (5.9) and setting $N_o = 1$, we get the following chain of equalities:

$$\bar{W} = WK^* = \sum_{k=0}^{n}\alpha_kv_kK^*$$
$$= \sum_{k=1}^{n}\alpha_kb_k\bar{v}_k + \sum_{k=0}^{n}\alpha_kN_kC_k$$
$$= \sum_{k=1}^{n}\alpha_kb_k\bar{v}_k + \sum_{k=0}^{n}\alpha_kN_k\left(\sum_{i=k+1}^{n}N_i\bar{v}_ip_{k,i} + p_{k,n+1}\right)$$
$$= \sum_{i=1}^{n}a_ib_i\bar{v}_i + \sum_{k=0}^{n-1}\alpha_kN_k\left(\sum_{i=k+1}^{n}N_i\bar{v}_ip_{k,i}\right) + \sum_{k=0}^{n}\alpha_kN_kp_{k,n+1}$$
$$= \sum_{i=1}^{n}\alpha_ib_i\bar{v}_i + \sum_{i=1}^{n}N_i\bar{v}_i\left(\sum_{k=0}^{i-1}\alpha_kN_kp_{k,i}\right) + \sum_{k=0}^{n}\alpha_kN_kp_{k,n+1}$$
$$= \sum_{i=1}^{n}\left(\alpha_ib_i + N_i\left(\sum_{k=0}^{i-1}\alpha_kN_kp_{k,i}\right)\right)\bar{v}_i + \sum_{k=0}^{n}\alpha_kN_kp_{k,n+1}$$
$$= \sum_{i=1}^{n}\beta_i\bar{v}_i + \beta_o,$$

which completes the proof. □

It is now clear how to truncate an ε-constructible part of a state space X.

ALGORITHM 5b. Let $\varepsilon \geqq 0$ be given. We set

$$k_c := \max\left\{k;\ \sum_{i=1}^{k-1}|\beta_i(\sigma)|^2 \leqq \varepsilon^2 \text{ for some } \sigma \in S_n\right\} \tag{5.10}$$

where the $\beta_i(\sigma)$ are as in (5.7), and set σ_c to be the maximizer of (5.10) (with the same precaution concerning nonuniqueness as in Algorithm 5a). Set

$$X_c := I_{\bar{u}}^{-1}\,\text{span}\,\{\bar{v}_{k_c}^{\sigma_c}, \cdots, \bar{v}_n^{\sigma_c}\},$$
$$y_c(0) := \int \sum_{i=k_c}^{n+1}\beta_i(\sigma)v_i^{\sigma}\,d\hat{\bar{u}}.$$

Then X_c is Markovian splitting and constructible for y_c. Moreover,

$$\|y(0) - y_c(0)\| \leqq \varepsilon.$$

To obtain an approximate model the general procedure is as follows.

ALGORITHM 5. Let $X \cong (u, \bar{u})$ and $\varepsilon \geqq 0$ be given. From X, using Algorithm 5a, derive the space $X_o \subset X$, and the process $\{y_o(t)\}$. Next, apply Algorithm 5b to X_o and $\{y_o(t)\}$, to get $X_{oc} \subset X_o$, and $\{y_{oc}(t)\}$. Then,

$$\|y(0) - y_{oc}(0)\| < 2\varepsilon.$$

An important question is whether the reduced model thus obtained is indeed minimal for the process $\{y_{oc}(0)\}$. We know that it is constructible (from the conjugate version of Theorem 5.2), but observability could have been destroyed a priori after the ε-unconstructible part was cut. The next theorem shows that this is not the case.

THEOREM 5.3. *Let y_{oc} and X_{oc} be, respectively, the process and the state space (splitting for y_{oc}) obtained through Algorithm 5. Then X_{oc} is minimal splitting for y_{oc}.*

Proof. All we need to show is the observability of X_{oc}, because constructibility follows from Theorem 5.2. Let W_{oc} be the spectral factor of the forward representation of y_{oc}, i.e.,

$$y(0) = \int W_{oc}\, d\hat{u}_{oc}$$

where $X_{oc} \cong (u_{oc}, \bar{u}_{oc})$. Let then $X_0 \cong (u_0, \bar{u}_0)$ be the observable space obtained with the first step of the procedure. Then, for any $\sigma \in S_{k_o}$, the coefficient $\alpha_{k_o}(\sigma)$ in the representation

$$W_o = \sum_{i=o}^{k_o} \alpha_i(\sigma) v_i^\sigma \tag{5.11}$$

is different from zero. We claim that

$$W_{oc} = \gamma + B_{\sigma_1} \cdots B_{\sigma_k} \left(\sum_{i=k_{oc}+1}^{k_o} \alpha_i(\sigma) v_i^\sigma \right) \tag{5.12}$$

where

$$\gamma = \sum_{i=0}^{k_{oc}} \left(\prod_{j=i+1}^{k_{oc}} (-b_{\sigma_j}) \right) \alpha_i N_i N_{k_{oc}}$$

and $N_o = 1$ for some $\sigma \in S_{k_{oc}}$. In fact, letting $K_o = W_o \bar{W}_o^{-1}$, we have

$$\begin{aligned} \bar{W}_o = W_o K_o^* &= \left(\sum_{i=0}^{k_o} \alpha_i(\sigma) v_i^\sigma \right) K_o^* \\ &= \left(\sum_{i=0}^{k_{oc}} \alpha_i(\sigma) v_i^\sigma \right) K_o^* + \left(\sum_{i=k_{oc}+1}^{k_o} \alpha_i(\sigma) v_i^\sigma \right) K_o^* \\ &= \sum_{i=1}^{k_{oc}} \beta_i(\sigma) \bar{v}_i^\sigma + \gamma C_{k_{oc}+1} + \left(\sum_{i=k_{oc}+1}^{k_o} \alpha_i(\sigma) v_i^\sigma \right) K_o^*. \end{aligned}$$

Since the second and the third added in the last term are orthogonal to span $\{\bar{v}_1^\sigma, \cdots, \bar{v}_i^\sigma\}$, when the ε-constructible part is truncated, we obtain exactly

$$\bar{W}_{oc} = \gamma_{k_{oc}+1} + \left(\sum_{i=k_{oc}+1}^{k_o} \alpha_i(\sigma) v_i^\sigma \right) K_0^*. \tag{5.13}$$

Multiplying (2.13) by $C_{k_{oc}+1}^*$ we get precisely (5.12). But the coefficients of the v_i are the same as the last $k_o - k_{oc}$ coefficients in (5.11), and observability then follows from that X_o.

Now we have the following important result about the consistency of Algorithm 5. In fact, we can easily show that the inclusion property, i.e., the very choice $\mathscr{Z} = \mathscr{Z}_I$ (cf. 3.4), implies consistency.

THEOREM 5.4. *Suppose an algorithm solves Problem* 1 *with* $\mathscr{Z} = \mathscr{Z}_I$. *Then this algorithm is consistent.*

Proof. Let $(\hat{y}, \hat{X})$ be a solution generated by our algorithm. Suppose there exists a minimal realization of y of dimension k equal to $\dim \hat{X}$. Since there exists in any state space X for y a minimal state space X_k contained in X, the minimal realization (y, X_k) of y belongs to $\mathscr{Z}_I$, and clearly with this choice the error (3.2) is minimized, since it is zero. So the error of $\hat{y}$ must also be zero, i.e., $\hat{y} = y$ almost surely, and hence $(y, \hat{X})$ is also a realization of y. Since its dimension equals that of X_k, it is minimal. Therefore the inclusion property implies consistency.

COROLLARY. *Algorithm* 5 *is consistent.*

We conclude with a remark on the infinite-dimensional case. One advantage of our procedure is that it extends very easily to the case when X has infinite dimension, and the structural function K is not rational. In fact, if K is a Blascke product, then the set (4.1) (which is now infinite-dimensional), forms a basis in $H(K)$, and hence W will have representation

$$W = \sum_{i=0}^{\infty} \alpha_i(\sigma) v_i^{\sigma}$$

with $\{\alpha_i(\sigma)\}_{i=0}^{\infty} \in l^2$. (This is because $\{v_i^{\sigma}\}$ is an orthonormal set.) That is, in the case when K is a Blascke product, the coefficients $\alpha_i(\sigma)$ will eventually become small. This does not happen with Hankel-norm approximation, where, to get a slower convergence of the singular values, the assumption $W \in H^{\infty} + C(\mathbb{T})$ is needed.

The general case, when K has a singular inner part, can be treated in a similar fashion. In fact, by Frostman's theorem [18], any inner function can be uniformly approximated by a Blascke product. Then, for any $\varepsilon > 0$, there exists a Blascke product B such that

$$\| W - E^{H(B)} W \| < \varepsilon,$$

and hence we can define the ε-unobservable part of X, etc. Again, this is not possible with Hankel-norm approximation.

Appendix: Proof of Theorem 3.1. A subspace X_1 of H^- is Markovian if and only if its image $\mathscr{X}_1$ under I_{u_-} has the form $\mathscr{X}_1 = Q(H^2 \ominus BH^2)$ with B_1, B_2 inner. Since $z\eta_0$ is outer, all we need to show is that $\mathscr{X}_1$ is not of the form $H^2 \ominus BH^2$, or equivalently (from Beurling's theorem), that $\mathscr{X}_1$, which is equal to $\operatorname{span}\{\eta_0, \cdots, \eta_{k-1}\}$, is not invariant for the left shift in $z^{-1}H^2$, which is to say that $\operatorname{span}\{z\eta_0, \cdots, z\eta_{k-1}\}$ is not left invariant in H^2. To this end, we consider the Hankel operator $\mathscr{H}_T$ from H^2 to $H^{2\perp}$:

$$\mathscr{H}_T := E^{H^{2\perp}} M_{T|H^2}.$$

We also define $T_- := E^{H^{2\perp}} T$. It is well known that $\mathscr{H}_T = \mathscr{H}_{T_-}$, and it is clear that

$$\mathscr{H} x_i = \sigma_i(\mathscr{H}_{T_-}) y_i \tag{A1}$$

where $x_i = z\eta_i$, $y_i = z\xi_i$, and $\sigma_i(\mathscr{H}_{T_-}) = \sigma_i(\mathscr{H}_T)$. Hence what we need to prove is that the space

$$\mathscr{H}_k := \operatorname{span}\{x_i;\ i = 0, \cdots, k-1\}$$

is not left invariant in H^2. To this end we need some technical lemmas.

LEMMA A1. *Let B be a Blaschke product, and let $H(B) := H^2 \ominus BH^2$. Then*

$$S(B) := E^{H(B)} M_{z^{-1}|H(B)} \tag{A2}$$

has no reducing subspace.

(A subspace X is reducing for an operator A if $AX \subset X$ and $AX^* \subset X^*$.)

Proof. The adjoint of $S(B)$ is easily seen to be $E^{H^2} M_{z|H(B)}$. Suppose Z_1 is reducing for $S(B)$. Then $A^* \mathscr{Z}_1 = E^{H^2} M_z \mathscr{Z}_1$ is contained in $\mathscr{Z}_1$. So $\mathscr{Z}_1$ is invariant for the left shift on H^2. For the same reason, $\mathscr{Z}_2 = X \ominus \mathscr{Z}_1$ is also invariant for the left shift U^*. But a subspace Z invariant for U^* has the form $Z = H^2 \ominus BH^2$ for some inner function Q (again from Beurling's theorem). Then

$$Z_1 = H^2 \ominus B_1 H^2 = H(B_1), \qquad Z_2 = H^2 \ominus B_2 H^2 = H(B_2)$$

with B_1, B_2 inner. We are going to show that $Z_1 \neq 0$ implies $Z_2 = 0$. Since Z_1 and Z_2 are orthogonal, Z_1 is contained in $Z_2^\perp = B_2 H^2$. Let $B_1 = q_1(z)/\tilde{q}_1(z)$. Then, $1/\tilde{q}_1(z) = B_2 f(z)$ for some $f \in H^2$; but $1/\tilde{q}_1(z)$ is outer, and hence $B_2 = 1$, i.e., $Z_2 = 0$.

From now on, B_{T_-} will denote the Blaschke product obtained with the poles of T_-.

LEMMA A2. *Let B_1 be an inner divisor of B_{T_-}. Then the operator*

$$\mathscr{H}_1 := \begin{cases} 0 & \text{on } H(B_1), \\ \mathscr{H} & \text{on } B_1 H^2 \end{cases} \tag{A3}$$

is Hankel if and only if the "prediction error" operator based on $H(B_1)$,

$$E_1 := (I - E^{H(B_1)}) M_{z^{-1}|H(B_1)},$$

has range contained in the kernel of H.

Proof. We recall that an operator H is Hankel if and only if $\mathscr{H} M_{z^{-1}} = E^{H^{2\perp}} M_z H$. In particular, since $H^2 = H^2(B_1) \oplus B_1 H^2$, $\mathscr{H}_1$ is Hankel if and only if

(i) $\mathscr{H}_1 M_{z^{-1}} f = E^{H^{2\perp}} M_{z^{-1}} f$ for $f \in B_1 H^2$;

(ii) $\mathscr{H}_1 M_{z^{-1}} f = 0$ for $f \in H^2 \ominus B_1 H^2$,

condition (i) is always satisfied. In fact, from the invariance of $B_1 H^2$ for $M_{z^{-1}}$, we have that $f \in B_1 H^2$ implies $M_{z^{-1}} f \in B_1 H^2$, and hence

$$\mathscr{H}_1 M_{z^{-1}} f = \mathscr{H} M_{z^{-1}} f = E^{H^{2\perp}} M_{z^{-1}} \mathscr{H} f = E^{H^{2\perp}} M_{z^{-1}} \mathscr{H}_1 f.$$

As for condition (ii), since $\mathscr{H}_1 = \mathscr{H} E^{B_1 H^2}$,

$$\mathscr{H}_1 M_{z^{-1}} f = \mathscr{H} E^{B_1 H^2} M_{z^{-1}} f = \mathscr{H}(I - E^{H(B_1)}) M_{z^{-1}} f.$$

Hence, condition (ii) is satisfied if and only if

$$\mathscr{H}(I - E^{H(B_1)}) M_{z^{-1}} = 0,$$

which is equivalent to $\mathscr{R}[(I - E^{H(B_1)}) M_{z-1}] \subset \ker \mathscr{H}$. □

LEMMA A3. *There is no left invariant subspace $H(B_1) \subset H(B)$ such that $\mathscr{H}_1$ is Hankel ($\mathscr{H}_1$ as in* (A3)).

Proof. In view of Lemma A2, if $\mathscr{H}_1$ is Hankel, then $(I - E^{H(B_1)}) M_{z^{-1}} H(B_1)$ is in the kernel BH^2 of $\mathscr{H}$. This means that $E^{H(B)}(I - E^{H(B_1)}) M_{z^{-1}} H(B_1) = 0$, which implies $S(B) H(B_1) \subset H(B_1)$. But this means that $H(B_1)$ is invariant for $S(B)$. Since it is already left invariant, it is reducing. But this contradicts Lemma A1.

We can now finish the proof of the theorem. Suppose that for some k, the space $\mathscr{X}_1 = \text{span}\{x_i;\ i = 0, \cdots, k-1\}$ is invariant for the adjoint shift. That is to say, $\mathscr{X}_1$ is equal to $(B_1 H^2)^\perp$ for some inner function $B_1 = q_1(z)/\tilde{q}_1(z)$ of degree k. The vector x_k admits in turn the representation $x_k = B_1(z)$, for some $f \in H^2$ of degree $n - k$ and outer

(cf. [1]). Express $T_-(z)$ as a sum of two rational functions, one in $(B_1H^2)^\perp$ and the other in B_1H^2:

$$T_-(z)=\frac{p_1(z)}{q_1(z)}+\frac{p_2(z)}{q_2(z)}=T_1(z)+T_2(z).$$

Since

$$T_1(z)x_k(z)=\frac{p_1(z)}{q_1(z)}f\in H^2$$

we get

$$E^{H^{2\perp}}T_1(z)=0. \tag{A4}$$

Now consider the Hankel operator

$$\mathscr{H}_2:=E^{H^{2\perp}}M_{T_2}.$$

In view of (A4), $\mathscr{H}w_2$ coincides with $\mathscr{H}$ on x_k, i.e.,

$$\mathscr{H}x_k=\mathscr{H}_2x_k.$$

That is to say, $\mathscr{H}$ and $\mathscr{H}_2$ coincide on the whole invariant subspace of H^2 generated by x_k, which is B_1H^2. This space contains x_i for $i\geqq k$ (because $x_i\perp\mathscr{X}_1$ for $i\geqq k$), i.e., $\mathscr{H}_2x_i=\sigma_i(\mathscr{H}_{T_-})y_i$, $i\geqq k$. Moreover, y_i is divisible by B_1^* (cf. [1]) for $i\geqq k$. Hence $\mathscr{H}_{T_1}^*y_i=0$ for $i\geqq k$, and

$$\mathscr{H}_{T_2}^*y_i=\mathscr{H}_{T_-}^*y_i=\sigma_i(\mathscr{H}_{T_-})x_i,\qquad i\geqq k,$$

which is to say that $\mathscr{X}_1\in\ker\mathscr{H}_2$ (Lemma A2), or equivalently,

$$\mathscr{H}_2:=\begin{cases}0 & \text{on } H(B_1),\\ \mathscr{H}_{T_-} & \text{on } B_1H^2.\end{cases}$$

But this, in view of Lemma A3, is a contradiction. □

Acknowledgments. We thank Professor Anders Lindquist for the patient and helpful advice he gave us during the long evenings of the Swedish winter, providing us with all possible kinds of suggestions that greatly contributed to the improvement of this paper. It is also a pleasure to thank Professors Chris Byrnes and Anthony Bloch for the encouragement and help they gave us during stimulating discussions at the Royal Institute of Technology in spring 1985; and the anonymous referee, whose kind and helpful comments largely contributed to the improvement of the paper.

REFERENCES

[1] V. M. ADAMJAN, D. A. AROV, AND M. G. KREIN, *Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-Takagí problem*, Math. USSR Sb., 15 (1971), pp. 31-73.

[2] H. AKAIKE, *Markovian representations of stochastic processes by canonical variables*, SIAM J. Control Optim., 13 (1975), pp. 162-173.

[3] A. BULTHEEL AND P. DEWILDE, *Orthogonal functions related to the Nevanlinna-Pick problem*, in Proc. Fourth Symposium on Mathematical Theory of Networks and Systems, N. Lavan, ed., Western Periodical, 1981.

[4] A. CAVAZZANA AND M. PAVON, *Principal component analysis for multivariable time series*, to appear.

[5] U. B. DESAI AND D. PAL, *A realization approach to stochastic model reduction and balanced stochastic realization*, in Proc. 16th Annual Conference on Information Science and Systems, Princeton, NJ, 1982.

[6] P. A. FUHRMANN, *Linear systems and operators in Hilbert space*, McGraw-Hill, New York, 1981.

[7] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bound*, Internat. J. Control, 39 (1984), pp. 1115–1193.
[8] K. HOFFMAN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.
[9] A. LINDQUIST AND M. PAVON, *On the structure of state space models for discrete-time stochastic vector processes*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 418–432.
[10] A. LINDQUIST AND G. PICCI, *Infinite dimensional stochastic realization of continuous-time stationary processes*, in Topics in Operator Theory, H. Dym and I. Gohberg, eds., Birkhäuser-Verlag, 1984.
[11] ———, *Realization theory for multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809–857.
[12] Y. A. ROZANOV, *Stationary Random Processes*, Holden-Day, San Francisco, 1967.
[13] A. GOMBANI, M. PAVON, AND B. COPPO, *On the Hankel-norm approximation of stationary increments processes*, in Proc. Seventh Symposium on Mathematical Theory of Networks and Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, New York, 1986.
[14] E. A. JONCKHEERE AND J. W. HELTON, *Power spectrum reduction by optimal Hankel-norm approximation of the phase of the outer spectral factor*, in Proc. American Control Conference, San Diego, CA, June 1984.
[15] B. A. FRANCIS AND G. ZAMES, *Design of H^∞-optimal multivariable feedback systems*, in Proc. 22nd IEEE Conference on Decision and Control, San Antonio, TX, 1983.
[16] A. GOMBANI AND M. PAVON, *On the Hankel-norm approximation of linear stochastic systems*, Systems Control Lett., 5 (1985), pp. 283–288.
[17] ———, *On approximate recursive prediction of stationary stochastic processes*, Stochastics, to appear.
[18] J. B. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 108-119, January 1989

007

DISTURBANCE DECOUPLING, (f, g)-INVARIANT AND CONTROLLABILITY SUBSPACES OF A CLASS OF HOMOGENEOUS POLYNOMIAL SYSTEMS*

WIJESURIYA P. DAYAWANSA† AND CLYDE F. MARTIN†

Abstract. This paper discusses control systems of the type

$$\dot{x} = f(x) + \sum_{i=1}^{m} b_i u_i + \sum_{j=1}^{k} d_j \omega_j, \qquad x \in \mathbb{R}^n$$

where $f(x)$ is a homogeneous polynomial vector field, b_i and d_i are constant vectors, and the output function is linear. It is shown that the well-known theory of disturbance decoupling of linear systems extends to this class in a very natural way. The resulting theory is far simpler than the general nonlinear theory. More important, all computations needed can be done using very simple algorithms, which require only a finite number of computations and use only methods from linear algebra.

Key words. polynomial systems, Lie algebras, decoupling, disturbance decoupling, (f, g)-invariance

AMS(MOS) subject classification. 93

1. Introduction.

We consider a system on $\mathbb{R}^n$ of the form

$$\dot{x} = f(x) + \sum_{i=1}^{m} b_i u_i + \sum_{j=1}^{k} d_j \omega_j, \tag{1}$$

$$y = Cx \tag{2}$$

where $f(x)$ is a vector field such that each entry is a homogeneous polynomial function of common degree p, b_i and d_j are constant vectors and C is an $n \times m$ matrix. The output is denoted by y, the inputs by u_i, and the disturbances by ω_j. Our primary purpose in this paper is to show that problems such as disturbance decoupling and controlling the state while keeping the output at zero can be easily solved in this class.

Nonlinear control systems of the type $\dot{x} = f(x) + g(x)u$ with f and g being polynomial vector fields previously have been the object of study by several authors (Brockett [3], Baillieul [1], [2], Jurdjevic and Kupka [10], and Bonnard and Tebbikh [5], [6]). Examples of type (1) include the very important problem of controlling the angular velocity of a rigid body. It is well known (see Baillieul [1] and Byrnes and Isidori [7], for example) that such a system can be described by

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{pmatrix} = \begin{pmatrix} a_1 x_2 x_3 \\ a_2 x_1 x_3 \\ a_3 x_1 x_2 \end{pmatrix} + Bu. \tag{3}$$

Therefore our theory developed here will solve the problem of disturbance decoupling of a rotating rigid body with linear output functions of angular velocities.

It was observed by Jurdjevic and Kupka [10] that, for systems of this type, the set of points that can be reached along trajectories with zero initial state is a subspace if the degree of f is odd. When the reachable subspace is $\mathbb{R}^n$, the state can be controlled from a given initial point to a desired final point in arbitrarily short time. This hints at the possibility of extending the geometric theory of linear control (see Wonham [15])

* Received by the editors September 21, 1987; accepted for publication (in revised form) April 11, 1988.
† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

to this class without having to go through the complicated computations needed in the local nonlinear geometric theory. In this paper we show that, indeed, the theory of disturbance decoupling extends naturally. Moreover, we show that when the degree of f is odd there exists a largest controllability subspace in a given subspace of $\mathbb{R}^n$ (see § 5 for definition). It is perhaps surprising that all computations can be done using linear algebraic methods and we give algorithms for this.

Notation 1.1.

B The $n \times m$ matrix of which the columns are $b_1, \cdots, b_m$.
D The $n \times k$ matrix of which the columns are $d_1, \cdots, d_k$.
$\mathcal{B}, \mathcal{D}$ The column spans of B and D, respectively.
$\mathcal{F}$ The collection of vector fields $\{f, b_1, \cdots, b_m\}$.
$\mathcal{F}^+$ The collection of vector fields $\{f, b_1, \cdots, b_m, d_1, \cdots, d_k\}$.
Let $\mathcal{X}$ be a given subset of vector fields on $\mathbb{R}^n$ and $x \in \mathbb{R}^n$ be arbitrary.
Lie $(\mathcal{X})$ The Lie algebra of vector fields generated by $\mathcal{X}$.
Lie $(\mathcal{X})(x)$ The subspace of $\mathbb{R}^n$ obtained by evaluating Lie $(\mathcal{X})$ at x.
$\text{Lie}_0(\mathcal{X})$ The subset of constant vector fields in Lie $(\mathcal{X})$.
Let $\mathcal{Y}$ be a subset of $\mathcal{X}$.
$C^q(\mathcal{X}; \mathcal{Y})$ The set of q tuples $\{Y_1, \cdots, Y_q\} \subset \mathcal{X}$ such that at least one $Y_i \in \mathcal{Y}$; q is an arbitrary positive integer.
$\mathcal{I}(\mathcal{X}; \mathcal{Y})$ Lie ideal generated by $\mathcal{Y}$ in Lie $(\mathcal{X})$.
$\mathcal{I}(\mathcal{X}; \mathcal{Y})(x)$ Evaluation of $\mathcal{I}(\mathcal{X}, \mathcal{Y})$ at x.
$\mathcal{I}_0(\mathcal{X}; \mathcal{Y})$ The subset of constant vector fields in $\mathcal{I}(\mathcal{X}; \mathcal{Y})$.

We will identify $\text{Lie}_0(\mathcal{X})$ and $\mathcal{I}_0(\mathcal{X}; \mathcal{Y})$, etc. with subspaces of $\mathbb{R}^n$ in a natural way. Also, when $\mathcal{X}$ or $\mathcal{Y} = \{d_1, \cdots, d_k\}$ or $\{b_1, \cdots, b_m\}$, we will denote them simply by D or B. We will not distinguish between constant vectors and constant vector fields of $\mathbb{R}^n$ unless the distinction is required to clarify the context.

The Lie algebraic computations needed in this paper can be done most conveniently by using multilinear algebra. In this regard, note that if $h: \mathbb{R}^n \to \mathbb{R}^m$ is a homogeneous polynomial function of degree p, then it is well known that we can associate a unique symmetric p-linear map

$$H: \prod_{i=1}^{p} \mathbb{R}^n \to \mathbb{R}^{na}$$

having the following property:

$$h(x) = H(x, x, \cdots, x).$$

For example, if $h: \mathbb{R}^2 \to \mathbb{R}$ is

$$h(x_1, x_2) = x_1^2 + x_1 x_2$$

then $H: \prod_{i=1}^{2} \mathbb{R}^2 \to \mathbb{R}$ is given by

$$H((x_1^1, x_2^1), (x_1^2, x_2^2)) = x_1^1 x_1^2 + \tfrac{1}{2}(x_1^1 x_2^2 + x_1^2 x_2^1).$$

Recall that the kth derivative at zero of h of degree p is a symmetric k-linear map $(D^k h(0)): \prod_{i=1}^{k} \mathbb{R}^n \to \mathbb{R}^m$. The map H is defined in general by

$$H = \frac{1}{p!}(D^p h(0)).$$

Now by using the Taylor series of h we obtain

$$h(x) = \sum_{k=0}^{p} \frac{1}{k!}(D^k h)(0)(x, x, \cdots, x).$$

But $(D^k h)(0) = 0$ if $k < p$. Hence,

$$\begin{aligned} h(x) &= \frac{1}{p!}((D^p h)(0))(x, x, \cdots, x) \\ &= H(x, x, \cdots, x), \end{aligned} \tag{4}$$

as desired.

The symmetric p-linear map associated to f will be denoted by F. The main purpose of introducing F is the following identity. Let $\nu^1, \cdots, \nu^p$ be arbitrary elements of $\mathbb{R}^n$. Since there is no ambiguity, we will denote the associated constant vector fields by the same symbols. Now $\mathrm{ad}_{\nu^1} \cdots \mathrm{ad}_{\nu^p}(f)$ is a constant vector field on $\mathbb{R}^n$ and an easy calculation shows that

$$\begin{aligned} \mathrm{ad}_{\nu^1} \cdots \mathrm{ad}_{\nu^p}(f) &= (D^p f)(0)(\nu^1, \cdots, \nu^p) \\ &= (p!)F(\nu^1, \cdots, \nu^p). \end{aligned} \tag{5}$$

2. Accessibility and reachability subspaces. The key observation regarding the class of systems (1) was made in Baillieul [2], Bonnard [6], and Jurdjevic and Kupka [10].

THEOREM 2.1 [2], [6], and [10]. *Assume that no disturbances are present.*

(i) *The set of points reachable from the origin is contained in* $\mathrm{Lie}_0(\mathscr{F})$ *and has nonempty interior in* $\mathrm{Lie}_0(\mathscr{F})$.

(ii) *If the degree p of* $f(x)$ *is odd, then all points in* $\mathrm{Lie}_0(\mathscr{F})$ *are reachable from the origin in arbitrarily short time.*

Proof. For a complete proof we refer the reader to Jurdjevic and Kupka [10]. We will give a proof of (i) to motivate the computations of the rest of the paper.

Since $\mathrm{Lie}(\mathscr{F})$ is obviously polynomial, it is also analytic and hence integrable (Nagano [11], Sussmann [14]). Now by Sussmann and Jurdjevic's results [13] on controllability, it follows that the reachable set $\mathscr{R}$ from the origin is contained in the leaf L of $\mathrm{Lie}(\mathscr{F})$ through the origin and contains a nonempty open subset of L. We only need to show that $L = \mathrm{Lie}_0(\mathscr{F})$.

First, $\mathrm{Lie}_0(\mathscr{F})$ is obviously contained in L and $\mathscr{B} \subset \mathrm{Lie}_0(\mathscr{F})$. Consider the following algorithm:

$$A^0 = \mathscr{B},$$

$$A^{i+1} = \mathrm{span}\{F(\nu^1, \cdots, \nu^p) \mid \{\nu^1, \cdots, \nu^p\} \subseteq A^i\} + A^i.$$

Since $F(\nu^1, \cdots, \nu^p) = \mathrm{ad}_{\nu^1} \cdots \mathrm{ad}_{\nu^p} f$, it follows that each A^i is contained in $\mathrm{Lie}_0(\mathscr{F})$. Since the A^i form an increasing sequence of subspaces, they converge to some subspace A in a finite number of steps. Now $A \subset \mathrm{Lie}_0(\mathscr{F})$. We claim that $\mathrm{Lie}_0(\mathscr{F}) \subset A$ also. Remembering that $\mathscr{R}$ contains a nonempty open subset of $\mathrm{Lie}_0(\mathscr{F})$, it is enough to prove that the vector field f is tangential to A at all points of A (for then the controlled trajectories can never leave A showing that $\mathscr{R} \subset A$). But by the construction of A, it follows that if $x \in A$, then $f(x) = F(x, x, \cdots, x) \in A$. Now

$$\mathscr{R} \subset A = \mathrm{Lie}_0(\mathscr{F}) \subset L$$

and $\mathscr{R}$ contains a nonempty open subset of $\mathrm{Lie}_0(\mathscr{F})$ and hence $A = \mathscr{R} = L$. □

Definition 2.2. The algorithm,

$$
\begin{aligned}
&A^0 = \mathcal{B},\\
&A^{i+1} = \operatorname{span}\{F(\nu^1, \cdots, \nu^p) \,|\, \{\nu^1, \cdots, \nu^p\} \subseteq A^i\} + A^i
\end{aligned} \tag{6}
$$

will be referred to as the accessibility subspace algorithm.

Remark 2.3. Observe that (6) needs only finitely many computations. Let $\{e^1, \cdots, e^j\}$ be a basis of A^i. Then

$$A^{i+1} = A^i + \operatorname{span}\{F(e^{i(1)}, e^{i(2)}, \cdots, e^{i(p)}) \,|\, 1 \leqq i(1) \leqq \cdots \leqq i(p) \leqq j\}.$$

Two equivalent ways to compute A^{i+1} are:

$$A^{i+1} = A^i + \operatorname{span}\{f(x) \,|\, x \in A\},$$

$$A^{i+1} = A^i + \operatorname{span}\{f(\pm e^{i(1)} \pm e^{i(2)} \pm \cdots \pm e^{i(p)}) \,|\, 1 \leqq i(1) \leqq \cdots \leqq i(p) \leqq j\}.$$

These apparently different ways of looking at the accessibility subspace algorithm will be useful in the sequel.

Assumption 2.4. For the rest of the paper we will assume that $\operatorname{Lie}_0(\mathcal{F}) = \mathbb{R}^n$.

Actually we do not need this assumption for any of our theorems, yet this assumption simplifies the notation and ideas considerably. All of our results will be valid for all trajectories of (1) (allowing disturbances) starting at the origin, for example, and hence is particular when $\operatorname{Lie}_0(\mathcal{F}^+) = \mathbb{R}^n$.

3. Conditions for the disturbance to not affect the output. For the moment let us ignore feedback and find conditions for ω to not affect the output. The important subspace needed here is $\mathcal{I}_0(\operatorname{Lie}(\mathcal{F}^+); D)$ (see Notation 1.1 for the definition). For convenience $\mathcal{I}$ and $\mathcal{I}_0$ stands for $\mathcal{I}(\operatorname{Lie}(\mathcal{F}^+); D)$ and $\mathcal{I}_0(\operatorname{Lie}(\mathcal{F}^+); D)$, respectively. The following lemma characterizes $\mathcal{I}_0$. Its proof was adapted from the proof of Lemma 5 in [10].

LEMMA 3.1. *$\mathcal{I}_0$ is the smallest subspace V of $\mathbb{R}^n$ that satisfies the property*

$$D \subset V,$$

$$F(\nu_1, \nu_2, \cdots, \nu_p) \in V \quad \textit{for all } \nu_1 \in V, \textit{ all } \nu_2, \cdots, \nu_p \in \mathbb{R}^n.$$

Proof. Remembering that $\operatorname{Lie}_0(\mathcal{F}^+) = \mathbb{R}^n$ by Assumption 2.4, it follows easily that $V \subset \mathcal{I}_0$. We need to prove the reverse inclusion.

Each element of $\mathcal{I}$ is a linear combination of elements in D along with vector fields of the form $\operatorname{ad}_{Y_1} \cdots \operatorname{ad}_{Y_{q-1}}(Y_q)$ for some q and $\{Y_1, \cdots, Y_q\} \in C^q(\mathcal{F}^+, D)$. Since $D \subset V$ and $B \subset \operatorname{Lie}_0(\mathcal{F}^+)$, every element of $\mathcal{I}$ is a linear combination of an element belonging to V along with elements of the form $\operatorname{ad}_{Y_1} \cdots \operatorname{ad}_{Y_{q-1}}(Y_q)$, where $\{Y_1, \cdots, Y_q\} \in C^q(f \cup \mathbb{R}^n; V)$. Moreover, each such summand is a homogeneous vector field. By induction we will prove that all constant vector fields of the form $\operatorname{ad}_{Y_1} \cdots \operatorname{ad}_{Y_{q-1}}(Y_q)$, where

$$\{Y_1, \cdots, Y_q\} \in C^q(f \cup \mathbb{R}^q; V)$$

belong to V. Thereby we prove that $\mathcal{I}_0 \subset V$. We abbreviate $C^q(f \cup \mathbb{R}^q; V)$ to C^q.

Now let $X = \operatorname{ad}_{Y_1} \cdots \operatorname{ad}_{Y_{q-1}}(Y_q)$ be constant and $\{Y_1, \cdots, Y_q\} \in C^q$. Let r be the number of indices j such that $Y_j = f$. Define the notion of length of X, $l(X)$, as the smallest r among all such representations of $X \cdot (l(X) = 0$ if $X \in V)$. We will prove our assertion by using induction on length.

If $l(X) = 0$ or 1, by definition $X \in V$. The induction hypothesis is: $l(X) \leqq s$ implies that $l(X) = 0 \cdot (s \in \mathbf{N})$.

Assume that the hypothesis is true for some s. Let $X = \mathrm{ad}_{Y_1} \cdots \mathrm{ad}_{Y_q}(Y_{q+1})$ be such that the number of indices j with $Y_j = f$ is equal to $s+1$. There is a smallest integer t, $t \leqq q$ such that $Y_{t+1} = f$. Hence $X = \mathrm{ad}_{Y_1} \cdots \mathrm{ad}_{Y_t} \mathrm{ad}_f(\tilde{Y})$, where $\tilde{Y} = \mathrm{ad}_{Y_{t+2}} \cdots \mathrm{ad}_{Y_q}(Y_{q+1})$. Using the Jacobi identity, we can rewrite X as a sum of terms of the form $[\mathrm{ad}_{Y_{\gamma_1}} \cdots \mathrm{ad}_{Y_{\gamma_i}}(f), \mathrm{ad}_{Y_{\mu_1}} \cdots \mathrm{ad}_{Y_{\mu_{t-i}}}(\tilde{Y})]$, where $\gamma_1 < \cdots < \gamma_i \leqq t$ and $\mu_1 < \cdots < \mu_{t-i} \leqq t$. We have only to consider nonzero constant summands, which fall into four types:

Type 1. $i = p$, $\mathrm{ad}_{Y_{\mu_1}} \cdots \mathrm{ad}_{Y_{\mu_{t-i}}}(\tilde{Y})$ is linear and $\{Y_{\gamma_1}, \cdots, Y_{\gamma_i}\} \cap V \neq \varnothing$.

Type 2. $i = p$, $\mathrm{ad}_{Y_{\mu_1}} \cdots \mathrm{ad}_{Y_{\mu_{t-i}}}(\tilde{Y})$ is linear, $\{Y_{\gamma_1}, \cdots, Y_{\gamma_i}\} \cap V = \varnothing$.

Type 3. $i = p-1$, $\mathrm{ad}_{Y_{\mu_1}} \cdots \mathrm{ad}_{Y_{\mu_{t-i}}}(\tilde{Y})$ is constant, $\{Y_{\gamma_1}, \cdots, Y_{\gamma_i}\} \cap V = \varnothing$.

Type 4. $i = p-1$, $\mathrm{ad}_{Y_{\mu_1}} \cdots \mathrm{ad}_{Y_{\mu_{t-i}}}(\tilde{Y})$ is constant, $\{Y_{\gamma_1}, \cdots, Y_{\gamma_i}\} \cap V \neq \varnothing$.

It is easily seen that each of these types has length less than or equal to s, and hence by the induction hypothesis has zero length. □

THEOREM 3.2. *The following conditions are equivalent*:

(1) *For all bounded measurable disturbance inputs and all initial states, the disturbance does not affect the output.*

(2) $\mathscr{I}_0(\mathrm{Lie}(\mathscr{F}^+); D)$ *is contained in the kernel of* C.

Proof. We abbreviate $\mathscr{I}(\mathrm{Lie}(\mathscr{F}^+); D)$ to $\mathscr{I}_0$.

(2⇒1). Let W be a complimentary subspace to $\mathscr{I}_0$ in $\mathbb{R}^n$.

Let us write $x(t) = \xi(t) + \eta(t)$, where $\xi(t) \in W$, $\eta(t) \in \mathscr{I}_0$.

Now (1) implies that

$$\dot{x}(t) = f(\xi(t)) + \sum_{r=0}^{p-1} \frac{1}{r!(p-r)!} \underbrace{F(\xi(t), \cdots, \xi(t)}_{r}, \eta(t), \cdots, \eta(t))$$

$$+ \sum_{i=1}^{m} b_i u_i(t) + \sum_{j=1}^{k} d_j \omega_j(t).$$

But by our characterization of $\mathscr{I}_0$ in Lemma 3.1, it follows that

$$F(\xi(t), \cdots, \xi(t), \eta(t), \cdots, \eta(t)) \in \mathscr{I}_0.$$

Since $D \subset \mathscr{I}_0$ we can also write

$$\dot{\xi}(t) = \tilde{f}(\xi(t)) + \sum_{i=1}^{m} \tilde{b}_i u_i(t)$$

where $\tilde{f}$ and $\tilde{b}_i$ are the projections of f and b_i onto W along $\mathscr{I}_0$. In particular, the dynamics $\xi(t)$ are independent of the disturbances and, since $y(t) = Cx(t) = C\xi(t)$, the disturbance does not affect the output.

(1⇒2). Since $\mathscr{I}$ is spanned by

$$\{\mathrm{ad}_{Y_1} \cdots \mathrm{ad}_{Y_{q-1}}(Y_q) \mid \{Y_1, \cdots, Y_q\} \in C^q(\mathscr{F}^+; D); q \in \mathbf{N}\},$$

it follows that $\mathscr{I}_0$ is spanned by such vector fields that are constant. But it is well known that if the disturbance does not affect the output, then $\mathrm{ad}_{Y_1} \cdots \mathrm{ad}_{Y_{q-1}}(Y_q)(x) \in \mathrm{Ker}(C)$ for all x, for all $\{Y_1, \cdots, Y_1\} \in C^q(\mathscr{F}^+; D)$, and for all q. (See Theorem 3.2 of Isidori [9].) This can be proved rather directly using the Fliess series [8]. □

Remark 3.3. Even the (⇒) part above can be proved using the Fliess series.

In view of the Theorem 3.2 we give the following definition.

Definition 3.4. The subspace $\mathscr{I}_0(\mathrm{Lie}(\mathscr{F}^+); D)$ will be called the disturbance confining subspace.

The disturbance confining subspace can be computed using the following algorithm and its proof follows at once from Lemma 3.1.

ALGORITHM 3.5. DISTURBANCE CONFINING SUBSPACE ALGORITHM.

$$V^0 = \mathscr{D},$$

$$V^{i+1} = V^i + \operatorname{span}\{F(v_1, v_2, \cdots, v_p) \mid v_1 \in V^i,\ v_j \in \mathbb{R}^n \text{ for } j > 1\}.$$

Since the subspaces V^i are increasing, the algorithm converges in finitely many steps. As was stated in Remark 2.3, all computations are finite.

4. Disturbance decoupling and (f, B)-invariant subspaces. In this section we will study the problem of constructing feedback such that the output functions will not be affected by the disturbance inputs. The class of feedback functions considered is of the form

$$u_i(t) = \alpha_i(x(t)) + v_i(t) \tag{7}$$

where $\alpha_i(x)$ is a homogeneous polynomial of degree p. Our reasons for selecting this class of feedback laws are the following:

(1) This class of feedback laws preserves the structure of the system. The slightly more general class $u_i(t) = \alpha_i(x(t)) + \sum_{j=1}^m \beta_{ij} v_j(t)$ with constant β_{ij}'s does not give more freedom, since the disturbance confining subspace algorithm does not depend on β_{ij}'s.

(2) We will show later that if the disturbances can be decoupled from the output with feedback of the form (7) with *analytic functions* $\alpha_i(x)$, then there also exists a *homogeneous polynomial feedback* $\alpha_i(x)$, which solves the problem.

Remark 4.1. Presently we do not know whether reason (2) is valid for the class of feedback of the form $u_i(t) = \alpha_i(x(t)) + \sum_{j=1}^m \beta_{ij}(x(t)) v_j(t)$ with analytic functions α_i and β_{ij}.

Except in the proof of Theorem 4.9, $\alpha_i(x)$ will denote a homogeneous polynomial of degree p, and $\alpha(x)$ will denote an m-tuple of such functions.

Let $\mathscr{F}_\alpha = \{f + B\alpha, b_1, \cdots, b_m\}$ and $\mathscr{F}_\alpha^+ = \{f + B\alpha, b_1, \cdots, b_m, d_1, \cdots, d_k\}$. We will abbreviate $\mathscr{I}_0(\text{Lie}(\mathscr{F}_\alpha^+); D)$ to $\mathscr{I}_0^\alpha$.

Now if there exists feedback α in (7) that decouples the disturbances ω from the output, then from Theorem 3.2, $\mathscr{I}_0^\alpha \subset \operatorname{Ker}(C)$. By the characterization of $\mathscr{I}_0$ in Lemma 3.1 we have

$$\operatorname{ad}_{v_1} \cdots \operatorname{ad}_{v_p}(f + B\alpha) \subset \mathscr{I}_0^\alpha \quad \text{for } v_1 \in \mathscr{I}_0^\alpha \text{ and } v_2, \cdots, v_p \in \mathbb{R}^n.$$

Hence $\operatorname{ad}_{v_1} \cdots \operatorname{ad}_{v_p}(f) \subset \mathscr{I}_0^\alpha + \mathscr{B}$.

Definition 4.2. A subspace V of $\mathbb{R}^n$ satisfying

$$F(v_1, \cdots, v_p) \in V + \mathscr{B} \quad \text{for all } v_1 \in V \text{ and all } v_j \in \mathbb{R}^n \text{ for } j > 1 \tag{8}$$

will be called an (f, B)-invariant subspace.

Note that (8) is equivalent to $f(x+v) = f(x) \bmod (V + \mathscr{B})$ for all $x \in \mathbb{R}^n$, and all $v \in V$.

It is clear that the set of (f, B)-invariant subspaces is closed under the addition of subspaces and thus has a maximal element V^* in the kernel of C.

Definition 4.3. V^* will be called the maximal (f, B)-invariant subspace contained in the kernel of C.

THEOREM 4.4. *There exists feedback of form* (7) *which decouples the output from the disturbance input if and only if*

$$\mathscr{D} \subset V^*.$$

Proof. ($\Rightarrow$) $\mathscr{I}_0^\alpha$ is an (f, B)-invariant subspace contained in the kernel of C and containing $\mathscr{D}$. But the maximality of V^* implies that $\mathscr{D} \subset \mathscr{I}_0^\alpha \subset V^*$.

($\Leftarrow$) We are going to construct feedback α such that

$$(f+B\alpha)(x+v)=(f+B\alpha)(x) \bmod V^* \quad \text{for all } x\in\mathbb{R}^n \text{ and all } v\in V^*.$$

Pick a basis $\{e_1,\cdots,e_n\}$ of $\mathbb{R}^n$ such that $\{e_1,\cdots,e_q\}$ is a basis of V^*. Now $\{e_{i_1}\cdots e_{i_p}\}_{1\leqq i_1\leqq\cdots\leqq i_p\leqq n}$ forms a basis of the symmetric covariant p-tensors S_p on $\mathbb{R}^n$. Define a linear function $\hat{\alpha}: S_p\to\mathbb{R}^m$ such that

$$(F+B\hat{\alpha})(e_{i_1}\cdots e_{i_p})\in V^* \quad \text{whenever } i_1\leqq q. \tag{9}$$

Now $\hat{\alpha}$ defines a unique homogeneous polynomial function α of degree p and this α satisfies our requirement.

Now Algorithm 3.5 for constructing $\mathscr{I}_0^\alpha$ shows at once that $\mathscr{I}_0^\alpha\subset V^*\subset \operatorname{Ker} C$. □

Remark 4.5. Part ($\Rightarrow$) was proved by Bonnard [6] for quadratic systems.

Remark 4.6. Note that this is similar to the corresponding linear theorem. In that case $p=1$ and our V^* coincides with the maximal (A, B)-invariant subspace in the kernel of C. In fact we can even extend the corresponding linear algorithm (see Wonham [15]) to our situation.

ALGORITHM 4.7. COMPUTATION OF THE MAXIMAL (f, B)-INVARIANT SUBSPACE IN THE KERNEL OF C. Let $\{e_1,\cdots,e_n\}$ be an arbitrary basis of $\mathbb{R}^n$. Define the set of linear maps $A_{i_1,\cdots,i_{p-1}}:\mathbb{R}^n\to\mathbb{R}^n$; $1\leqq i_1\leqq\cdots\leqq i_{p-1}\leqq n$:

$$A_{i_1}\cdots i_{p-1}(x)=F(e_{i_1},\cdots,e_{i_{p-1}},x).$$

The collection of all such maps will be denoted by $\{A_i\}_{i\in I}$, where

$$I=\{(i_1,\cdots,i_{p-1})\,|\,1\leqq i_1\leqq\cdots\leqq i_{p-1}\leqq n\}.$$

Algorithm.

$$V^0=\operatorname{Ker}(C), \tag{10}$$

$$V^{i+1}=V^i\cap\Big(\bigcap_{i\in I}A_i^{-1}(V^i+B)\Big). \tag{11}$$

Clearly the sequence of subspaces is decreasing, and hence it converges to some subspace V^+. Moreover, $V^*\subset V^0$ by definition. Now suppose that $V^*\subset V^i$ for some i. Then it is clear from (11) that $V^*\subset V^{i+1}$. This proves by induction that $V^*\subset V^+$ and the maximality of V^*. The fact that V^+ satisfies (8) implies that $V^*=V^+$.

Remark 4.8. When $d_i\in V^*$, the disturbance decoupling feedback $\alpha(x)$ can be easily computed using (9).

We will show that the class of feedback functions we have been considering is fairly general.

THEOREM 4.9. *Suppose that there exists analytic feedback $u_i(t)=\alpha_i(x(t))+v_i(t)$ defined in a neighborhood of the origin that renders the closed-loop system disturbance decoupled. Then there exist feedback functions $\tilde{\alpha}_i(x)$ that are homogeneous of degree p that decouple the closed-loop system on $\mathbb{R}^n$.*

Proof. We will denote the m-tuple $\{\alpha_1,\cdots,\alpha_m\}$ by α and write $\alpha(x)=\sum_{j=0}^{\infty}\alpha^j(x)$ using the Taylor series. ($\alpha^j(x)$ is homogeneous of degree j.) Define $\tilde{\alpha}=\alpha^p$. Let f^α and $f^{\tilde{\alpha}}$ denote $f+B\alpha$ and $f+B\tilde{\alpha}$, respectively. Let $\mathscr{F}_\alpha^+=\{f^\alpha, B, D\}$ and $\mathscr{F}_{\tilde{\alpha}}^+=\{f^{\tilde{\alpha}}, B, D\}$. Similarly, we abbreviate $C^q(\mathscr{F}_\alpha^+; D)$ and $C^q(\mathscr{F}_{\tilde{\alpha}}^+; D)$ to C_α^q and $C_{\tilde{\alpha}}^q$, respectively. We also abbreviate $\mathscr{I}(\operatorname{Lie}(\mathscr{F}_\alpha^+); D)$, $\mathscr{I}(\operatorname{Lie}(\mathscr{F}_{\tilde{\alpha}}^+); D)$; $\mathscr{I}_0(\operatorname{Lie}(\mathscr{F}_\alpha^+); D)$ and $\mathscr{I}_0(\operatorname{Lie}(\mathscr{F}_{\tilde{\alpha}}^+); D)$ to $\mathscr{I}^\alpha$, $\mathscr{I}^{\tilde{\alpha}}$, $\mathscr{I}_0^\alpha$ and $\mathscr{I}_0^{\tilde{\alpha}}$, respectively.

Now $\mathscr{I}^\alpha$ is spanned by $\{\operatorname{ad}_{Y_1}\cdots\operatorname{ad}_{Y_{q-1}}(Y_q)\,|\,\{Y_1,\cdots,Y_q\}\in C_\alpha^q, q\geqq 1\}$. For each $\{Y_1,\cdots,Y_q\}\in C_\alpha^q$ we will denote by $\{\tilde{Y}_1,\cdots,\tilde{Y}_q\}$ the q-tuple in $C_{\tilde{\alpha}}^q$ obtained by

replacing feedback α by $\tilde{\alpha}$. Now suppose that $\{Y_1, \cdots, Y_q\} \in C_\alpha^q$ is such that $\operatorname{ad}_{\tilde{Y}_1} \cdots \operatorname{ad}_{\tilde{Y}_{q-1}}(\tilde{Y}_q) \in \mathcal{I}_0^{\tilde{\alpha}}$. Then $(\operatorname{ad}_{Y_1} \cdots \operatorname{ad}_{Y_{q-1}}(Y_l))(0) = \operatorname{ad}_{\tilde{Y}_1} \cdots \operatorname{ad}_{\tilde{Y}_{q-1}}(\tilde{Y}_q)$ and from this we conclude that $\mathcal{I}_0^{\tilde{\alpha}} \subset \mathcal{I}_0^{\alpha}$. But since $\mathcal{I}^{\alpha} \subset \operatorname{Ker} C$, it follows that $\mathcal{I}_0^{\tilde{\alpha}} \subset \operatorname{Ker} C$. We conclude by Theorem 3.2 that the feedback $u = \tilde{\alpha} + v$ decouples the disturbance and output. □

5. Reachability subspaces. Consider the following problem. We are given

$$\dot{x} = f(x) + Bu(t), \tag{12}$$

$$y = Cx \tag{13}$$

where f is a homogeneous polynomial of degree p, and B and C are $n \times m$ and $l \times n$ matrices. Can we partition the inputs, i.e., find an $m \times m$ nonsingular matrix $G = [G_{1n \times m_1} | G_{2n \times m_2}]$ and feedback

$$u = \alpha(x) + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \begin{matrix} m_1 \times 1 \\ m_2 \times 1 \end{matrix}$$

where $\alpha(x)$ is a homogeneous polynomial of degree p such that when we write the closed-loop system as

$$\dot{x} = (f + \alpha)(x) + BG_1 v_1(t) + BG_2 v_2(t) \tag{14}$$

the inputs $v_1(t)$ do not affect the output y? In this section we give necessary and sufficient conditions for the existence of a solution to this problem, and show that when there is a solution, it can be found such that the set of points that can be reached by $v_1(t)$ is maximized. This problem is important in controlling a system while keeping the output at a constant value or when decoupling the effect of the inputs on the outputs.

Now let us consider the system (12), (13). Suppose that $v_2(t) = 0$ and $x(0) = 0$. Then it follows from Theorem 2.1 that the set of points that can be reached along measurable or piecewise constant control trajectories is contained in the subspace of constant vector fields in the Lie algebra generated by $\{f + B\alpha, BG_1\}$, where BG_1 denotes the vector fields given by the columns of BG_1. We will denote this subspace by

$$\langle f + B\alpha \,|\, \operatorname{Im}(BG_1) \rangle$$

and call it the *accessibility subspace* of $\{f + B\alpha, \operatorname{Im}(BG_1)\}$. Theorem 2.1 states that if p is odd, this space is the space of points that can be reached from the origin. In this case we can also call them controllability subspaces. Our theory here will be the counterpart of the theory of (A, B)-controllability subspaces of a linear system. Our proofs are generalizations of the corresponding proofs for (A, B)-controllability subspaces, as can be found in Wonham [15]. The subspace $\langle f + B\alpha \,|\, \operatorname{Im}(BG_1) \rangle$ can be computed using the accessibility subspace algorithm (6). We advise the reader to keep this algorithm in mind since it will be helpful in understanding the proofs of this section. As before we will assume that $\langle f \,|\, \operatorname{Im}(B) \rangle = \mathbb{R}^n$ for simplicity of exposition. We denote by $\mathcal{B}$, $\mathcal{B}_1$, $\mathcal{B}_2$, etc. the images of B, B_1, and B_2. When we refer to feedback α, we mean that α is a homogeneous polynomial function of degree p.

Notation 5.1. We define (f, B)-invariant subspaces as in § 4. The collection of all (f, B)-invariant subspaces contained in a given subspace $\mathcal{K}$ will be denoted by $S(f, B; \mathcal{K})$. As we saw in the proof of Theorem 4.4, for all $V \in S(f, B; \mathcal{K})$ there exists feedback α such that

$$\operatorname{ad}_{v_1} \cdots \operatorname{ad}_{v_{p-1}} \operatorname{ad}_x (f + B\alpha) \in V \quad \text{for all } v_1, \cdots, v_{p-1} \in \mathbb{R}^n \quad \text{for all } x \in V.$$

The set of all such feedback will be denoted by $\Lambda(V)$.

Definition 5.2. A subspace $\mathscr{R}$ of $\mathbb{R}^n$ is called an (f, B)-accessibility subspace if $\mathscr{R} = \langle f + B\alpha \,|\, \mathrm{Im}\,(BG_1)\rangle$ for some feedback α and G_1. The collection of all (f, B)-accessibility subspaces contained in a given subspace $\mathscr{K}$ will be denoted by $\mathscr{A}(f, B; \mathscr{K})$.

Let $\mathscr{K}$ be the kernel of C. If there is a solution to our problem, then according to Theorem 4.4, $\mathrm{Im}\,(BG_1)$ should be contained in V^*, the largest element of $S(f, B; \mathscr{K})$. Conversely, if there exist G_1 such that $\mathrm{Im}\,(BG_1) \subset V^*$, then we can find feedback $\alpha + G_1 v_1 + G_2 v_2$ such that the input v_1 does not affect the output. This completely solves the problem we posed earlier. However, we proceed further and ask whether G_1 can be found such that $\langle f + B\alpha \,|\, \mathrm{Im}\,(BG_1)\rangle$ is also maximal. In a sense, we are talking about maximizing our ability to control the state while keeping $y(t) \equiv 0$. We prove the existence of such a maximal (f, B)-accessibility subspace and give an algorithm to compute α, G_1, and G_2. We will proceed via a series of lemmas. We caution that if R is an (f, B)-accessibility subspace, then it may not be an (f, B)-invariant subspace. However, $\mathrm{ad}_{v_1} \cdots \mathrm{ad}_{v_p}(f) \in R + B$ for all $v_i \in R$. Hence it makes sense to define $\Lambda(R)$, the set of feedback functions α such that $(f + B\alpha)(V) \subset V$.

We reiterate that the basic ideas of the proofs are due to Wonham [15].

LEMMA 5.3. *Let G_1 be arbitrary and define $\mathscr{R} = \langle f \,|\, \mathrm{Im}\,(BG_1)\rangle$. Then $R = \langle f \,|\, \mathscr{B} \cap \mathscr{R}\rangle$. Conversely, if $R = \langle f \,|\, \mathscr{B} \cap \mathscr{R}\rangle$, then there exist G_1 such that $R = \langle f \,|\, \mathrm{Im}\,(BG_1)\rangle$.*

Proof. The second assertion is obvious. So we turn to the first one. Since $\mathrm{Im}\,(BG_1) \subset R$, the accessibility algorithm implies that $R \subset \langle f \,|\, \mathscr{B} \cap \mathscr{R}\rangle$. Conversely, $\mathscr{B} \cap \mathscr{R} \subset R$ and $f(R) \subset R$. Hence $\langle f \,|\, \mathscr{B} \cap \mathscr{R}\rangle \subset \mathscr{R}$. □

We obtain the following corollary.

COROLLARY 5.4. *$\mathscr{R} \in \mathscr{A}(f, B; \mathscr{K})$ if and only if there exist α such that*

$$R = \langle f + B\alpha \,|\, \mathscr{B} \cap R\rangle.$$

LEMMA 5.5. *If $\mathscr{R} \in \mathscr{A}(f, B; \mathscr{K})$, then $\mathscr{R} = \langle f + B\alpha \,|\, \mathscr{B} \cap \mathscr{R}\rangle$ for all $\alpha \in \Lambda(R)$.*

Proof. If $\mathscr{R} \in \mathscr{A}(f, B; \mathscr{K})$, then by definition there exist α_0 such that $R = \langle f + B\alpha_0 \,|\, \mathscr{B} \cap \mathscr{R}\rangle$. Let $\alpha_1 \in \Lambda(R)$ be arbitrary and let

$$\mathscr{R}_1 = \langle f + B\alpha_1 \,|\, \mathscr{B} \cap \mathscr{R}\rangle.$$

Since $(f + B\alpha_1)\mathscr{R} \subset \mathscr{R}$, it follows easily that $\mathscr{R}_1 \subset \mathscr{R}$. To prove the converse we use the accessibility subspace algorithm.

Let $V_0 = \mathscr{B} \cap \mathscr{R}$ and

$$V^{i+1} = \mathrm{span}\,\{(f + B\alpha_0)(v) \,|\, v \in V^i\} + V^i.$$

The induction hypothesis is

$$V^i \subset \mathscr{R} \quad \text{for } i \leqq k.$$

Clearly this is true for $k = 0$. Suppose that it is true for some k. Let $x \in V^k$. Then

$$(f + B\alpha_0)(x) = (f + B\alpha_1)(x) + B(\alpha_0(x) - \alpha_1(x)).$$

By the accessibility subspace algorithm, $(f + B\alpha_1)(x) \in \mathscr{R}_1$. Also, $B(\alpha_0(x) - \alpha_1(x)) \in B$ and

$$B(\alpha_0(x) - \alpha_1(x)) = (f + B\alpha_0)(x) - (f + B\alpha_1)(x) \in \mathscr{R}.$$

Hence $(f + B\alpha_0)(x) \in R_1 + (\mathscr{B} \cap \mathscr{R}) \subset R_1$, implying that $V^{k+1} \subset R_1$. By induction, $\mathscr{R} = \mathscr{R}_1$. □

LEMMA 5.6. *Let* $V \in S(f, B; \mathcal{K})$, $\mathcal{B}_0 \subset \mathcal{B} \cap \mathcal{V}$ *and* $\alpha_0 \in \Lambda(V)$ *and* $B(\alpha - \alpha_0)(V) \subset B_0$. *Then* $\mathcal{R} = \langle f + B\alpha \mid \mathcal{B}_0 \rangle$.

Proof. Let $\alpha \in \Lambda(V)$ and let $\mathcal{R}_1 = \langle f + B\alpha \mid \mathcal{B}_0 \rangle$. Define

$$V_0 = \mathcal{B}_0,$$

$$V^{i+1} = \text{span}\{(f + B\alpha_0)(v) \mid v \in V^i\} + V^i.$$

Then $V_0 = \mathcal{B}_0 \subset \mathcal{R}_1$. Suppose that $V^i \subset \mathcal{R}_1$ for $i \leqq k$ for some k. Let $x \in V^k$. Then, since obviously $\mathcal{R} \subset V$, it follows that

$$B(\alpha - \alpha_0)(x) \in V \cap \mathcal{B} = \mathcal{B}_0 \subset \mathcal{R}_1.$$

Hence

$$(f + B\alpha_0)(x) = (f + B\alpha_1)(x) + B(\alpha - \alpha_0)(x) \in \mathcal{R}_1.$$

By induction, $\langle f + B\alpha_0 \mid \mathcal{B}_0 \rangle \subset \mathcal{R}_1$. The inverse inclusion follows by the symmetry of the argument. □

LEMMA 5.7. *Let* $V \in S(f, B, \mathcal{K})$ *and let* $\mathcal{R} \subset V$ *and let* α_0 *be such that* $(f + B\alpha_0)(\mathcal{R}) \subset \mathcal{R}$. *Then there exist* $\alpha \in \Lambda(V) \cap \Lambda(R)$ *such that* $\alpha|_{\mathcal{R}} = \alpha_0|_{\mathcal{R}}$.

Proof. The proof follows by arguing as in the construction of α in the proof of Theorem 4.4. □

Now we are ready to state our maximality theorem.

THEOREM 5.8. *Let* $\mathcal{K}$ *be a given subspace and let* V^* *be the maximal* (f, B)*-invariant subspace contained in* $\mathcal{K}$. *Then there exists a maximal* (f, B)*-accessibility subspace* $\mathcal{R}^*$ *contained in* V^* *and* $\mathcal{R}^* = \langle f + B\alpha \mid V^* \cap \mathcal{B} \rangle$ *for arbitrary* $\alpha \in \Lambda(V^*)$.

Proof. Let $\alpha \in \Lambda(V^*)$ and let $\mathcal{R} = \langle f + B\alpha \mid \mathcal{B} \cap V^* \rangle$. Clearly $\mathcal{R} \subset V^*$. Let $\mathcal{R}_0 = \langle f + B\alpha_0 \mid \mathcal{B} \cap \mathcal{R}_0 \rangle$ be an arbitrary (f, B)-accessibility subspace in V^*. Pick $\alpha_1 \in \Lambda(R_0) \cap \Lambda(V^*)$ such that $\alpha_1 | R_0 = \alpha_0 R_0$. Now if $x \in V^*$, then

$$B(\alpha_1 - \alpha_0)(x) = (f + B\alpha_1)(x) - (f + B\alpha_0)(x) \in V^*.$$

Hence $B(\alpha - \alpha_0)(x) \in V^* \cap \mathcal{B}$ and now by Lemma 5.6,

$$\mathcal{R}_0 = \langle f + B\alpha_1 \mid \mathcal{B} \cap \mathcal{R}_0 \rangle \subset \langle f + B\alpha \mid \mathcal{B} \cap V^* \rangle = \mathcal{R}^*. \qquad \square$$

We also obtain the following important corollary.

COROLLARY 5.9. $\Lambda(V^*) \subset \Lambda(R^*)$.

Remark 5.10. Theorem 5.8 together with the accessibility subspace algorithm and the (f, B)-invariant subspace algorithm provides a means for computing the largest accessibility subspace.

6. Decoupling problem. We consider the system

$$\dot{x} = f(x) + Bu, \tag{15}$$

$$y_i = C_i x, \qquad i = 1, \cdots, r \tag{16}$$

where $f(x)$ (as before) is a homogeneous polynomial vector field of degree p; B is an $n \times m$ matrix; C_i is an $l_i \times n$ matrix. We propose a question. Can we find feedback

$$u = \alpha(x) + [G_1 | \cdots | G_r] \begin{bmatrix} v_1 \\ \vdots \\ v_r \end{bmatrix}$$

where $\alpha(x)$ is a homogeneous polynomial vector of degree p and $G_1, \cdots, G_r$ are constant matrices with the following properties:

(a) v_i does not affect y_j if $i \neq j$;

(b) The output y_i can be controlled to any desired value by v_i. We will assume that the subspaces $\{\text{Image}\,(C_i)\}_{i=1}^n$ are independent, for otherwise this problem does not make sense.

In view of Theorem 2.1, we will henceforth assume that p is odd. In view of the machinery we have at our disposal we can study this problem using (f, B)-accessibility subspaces just as in the linear case.

Notation 6.1. Let $\mathcal{K}$ be a subspace. Let $V^*(\mathcal{K})$ denote the largest (f, B)-invariant subspace in $\mathcal{K}$ and let $\mathcal{R}^*(\mathcal{K})$ denote the largest (f, B)-accessibility subspace contained in $V^*(\mathcal{K})$. If V is an (f, B)-invariant subspace, then $\Lambda(V)$ denotes the set of feedback functions α such that $(f+B\alpha)(x+v) = (f+B\alpha)(x) \bmod V$ for all $x \in \mathbb{R}^n$ and all $v \in V$.

THEOREM 6.2. *Suppose that there exists α such that*

(1) $\alpha \in \Lambda(V^*(\cap_{i\neq j} \ker(C_i)))$ *for* $j = 1, \cdots, r$,

(2) $\dim(C_j(\mathcal{R}^*(\cap_{i\neq j} \ker(C_i)))) = \text{rank}\, C_j$.

Then the decoupling problem has a solution.

Proof. Define matrices G_j such that

$$\text{Image}\,(BG_j) = \mathcal{R}^*\Big(\bigcap_{i\neq j} \ker(C_i)\Big) \cap \mathcal{B}, \qquad j = 1, \cdots, r.$$

Then by Theorem 5.8,

$$\mathcal{R}^*\Big(\bigcap_{i\neq J} \ker(C_i)\Big) = \langle f+B\alpha \,|\, \text{Image}\,(BG_j)\rangle, \qquad j = 1, \cdots, r.$$

Therefore by two, the input v_j completely controls the output y_j.

Moreover, our theory of disturbance decoupling (Theorem 4.4 and its proof) shows that y_j is not affected by v_i if $i \neq j$. □

In the previous theorem, having to verify the existence of α is cumbersome. If we want to decouple the state, a much nicer set of necessary and sufficient conditions can be stated. Let us denote by C the matrix $[C_1^T, \cdots, C_r^T]^T$.

THEOREM 6.3. *If* $\text{rank}\,(C) = n$, *then the decoupling problem is solvable if and only if*

$$\ker C_j + \mathcal{R}^*\Big(\bigcap_{i\neq j} \ker C_j\Big) = \mathbb{R}^n \quad \textit{for all } j.$$

Proof. The proof of necessity is trivial.

Sufficiency. Clearly our condition implies condition (2) of Theorem 6.2. Let $\mathcal{K}_j$ denote $\cap_{i\neq j} \ker C_i$. Then $\mathcal{K}_j \cap (\sum_{i\neq j} \mathcal{K}_i) = 0$, for all j. This follows at once since $\text{rank}\,(C) = n$. Therefore $V^*(\mathcal{K}_j) \cap (\sum_{k\neq j} (V^*(\mathcal{K}_j))) = 0$ for all j. Now the construction of α described in the proof of Theorem 4.4 easily shows that there exists some $\alpha \in \Lambda(V^*(\mathcal{K}))$, $j = 1, \cdots, r$. The sufficiency follows by Theorem 6.2. □

7. Concluding remarks. The primary purpose of this paper has been to show that the geometric theory of decoupling of linear control systems extends almost directly to the class of nonlinear systems given by (1) and (2). Along with Jurdjevic's and Kupka's theorem [10] on controllability quoted in Theorem 2.1, it almost seems to imply that the linear control theory we know of is actually an "odd degree polynomial" control theory. This merits investigation of the relationship between stabilizability and controllability. Unfortunately, controllability does not imply stabilizability even for this class.

Example 7.1. Consider the following system:

$$\dot{x}_1 = x_2^2 x_3,$$

$$\dot{x}_2 = x_3^3,$$

$$\dot{x}_3 = u.$$

This system is controllable. However, according to Brockett [4] and Byrnes and Isidori [7] a necessary condition for asymptotic stabilizability of this system is the existence of a function $\alpha(x_1, x_2, x_3)$ that renders the equation

$$\begin{bmatrix} x_2^2 x_3 \\ x_3^3 \\ \alpha(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

solvable for x_1, x_2, x_3 whenever a_1, a_2, a_3 are arbitrarily specified small real numbers. However, it is seen that when $a_2 = 0$, then necessarily $a_1 = 0$ also for solvability, showing that the system is not asymptotically stabilizable.

We do not know of any reasonably simple extra conditions that would ensure stabilizability.

REFERENCES

[1] J. BAILLIEUL, *The geometry of homogeneous polynomial dynamical systems*, Nonlinear Anal. Theory Methods Appl., 4 (1980), pp. 879-900.

[2] ———, *Controllability and observability of polynomial dynamical systems*, Nonlinear Anal. Theory Methods Appl., 5 (1981), pp. 543-552.

[3] R. W. BROCKETT, *Lie algebras and Lie groups in control theory*, in Geometric Methods in Control Theory, D. Q. Mayne and R. W. Brockett, eds., Reidel, Dordrecht, the Netherlands, 1973.

[4] ———, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millmann, and H. Sussmann, eds., Birkhäuser, Boston, 1983.

[5] B. BONNARD AND H. TEBBIKH, *Quadratic control systems*, preprint.

[6] B. BONNARD, *Controlabilité et observabilité d'une classe de systèmes non-linéaires*, Note, Laboratoire d'Automatique de Grenoble, Grenoble, 1981.

[7] C. I. BYRNES AND A. ISIDORI, *On the attitude stabilization of rigid spacecraft*, preprint.

[8] M. FLIESS, *Fonctionnelles causales non lineaires et indeterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3-40.

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

[10] V. JURDJEVIC AND I. KUPKA, *Polynomial control systems*, Math. Ann., 272 (1985), pp. 361-368.

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

[12] E. D. SONTAG AND Y. ROUCHALEAU, *On discrete time polynomial systems*, Nonlinear Anal. Theory Methods Appl., 1 (1976), pp. 55-64.

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

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

[15] M. W. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, Berlin, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 120–147, January 1989

008

THE STRUCTURE OF SMALL-TIME REACHABLE SETS IN LOW DIMENSIONS*

ARTHUR J. KRENER† AND HEINZ SCHÄTTLER‡

Abstract. This paper outlines a general method to determine the geometric structure of small-time reachable sets for a single-input control system with a bounded linear control. The authors' analysis relies on free nilpotent systems as a guide, and hence their techniques only apply to nondegenerate situations. The paper illustrates the effectiveness of the method in low dimensions. Among other results is given a precise description of the small-time reachable set for a system $\dot{x} = f(x) + g(x)u$, $|u| \leqq 1$ in dimension four, under the generic assumption that the constant controls $u \equiv +1$ and $u \equiv -1$ are not singular. As a corollary, a local synthesis is obtained in dimension three for the time-optimal control problem under the analogous generic condition.

Key words. nonlinear systems, nilpotent approximation, reachable sets, bang-bang trajectories, singular arcs

AMS(MOS) subject classifications. 49B10, 93B10

1. Introduction.

In this paper we study the qualitative structure of small-time reachable sets in low dimensions for a single-input system with a bounded linear control. More precisely, we consider a system of the form

$$\Sigma: \dot{x} = f(x) + g(x)u, \quad |u| \leqq 1, \quad x \in \mathbb{R}^n \tag{1}$$

where f and g are smooth (C^∞) or analytic vector fields and admissible controls are measurable functions with values in $[-1, 1]$ almost everywhere. A trajectory of the system corresponding to a control $u(\cdot)$ is an absolutely continuous curve $x(\cdot)$ such that $\dot{x}(t) = f(x(t)) + g(x(t))u(t)$ almost everywhere. We say a point q is reachable from a point p within time T if and only if there exists a trajectory $x(\cdot)$ defined on an interval $[0, t]$, $t \leqq T$, such that $x(0) = p$ and $x(t) = q$. The set of all such points q is denoted by Reach $(p, \leqq T)$; Reach (p, T) denotes the set of points that are reachable exactly at time T. The reachable set from p, Reach (p), is the set of all points that are reachable from p within some time T.

Reachable sets play an important role in control theory. If a system can be stabilized to a given point by a feedback control law, then that point must be in the reachable set of every other point. In optimal control problems, if the cost is added as another coordinate, then the optimal trajectories must lie in the boundary of the set of reachable points. For this reason the Pontryagin Maximum Principle plays an important role in studying the boundaries of reachable sets.

The problem of describing a reachable set and the extremal trajectories that generate its boundary is closely related to the problem of regular synthesis in the sense of Boltyansky [1] and others [5], [18]. While the problem has been studied extensively for many years, only a few examples of regular syntheses have been described, for instance, [24]. Even in low dimensions, the reachable set of a general control system can be extremely complicated.

* Received by the editors June 22, 1987; accepted for publication (in revised form) April 11, 1988. This research was partly supported by National Science Foundation grant DMS-8601635 and Air Force Office of Scientific Research grant 85-0267.

† University of California at Davis, Department of Mathematics, Davis, California 95616.

‡ Washington University, Department of Systems Science and Mathematics, St. Louis, Missouri 63130.

We shall attempt to avoid this difficulty by considering only "nondegenerate" systems. By a nondegenerate system we mean one where (i) f, g, and the low-order Lie brackets of f and g span as many dimensions as is possible given the dimensions of the state space; and where (ii) no nontrivial equality relations hold between those vector fields (for instance, if n is the space dimension, then any relation saying that n vector fields are dependent at a point is considered a nontrivial equality relation, whereas a relation that simply expresses the fact that a vector field can be written in terms of a basis is considered trivial).

This is in the spirit of Lobry [14], who described the small-time reachable set of (1) in dimension three under the assumption that f, g, and $[f, g]$ are linearly independent. The method described below is an attempt to extend Lobry's result to higher dimensions. As will be seen, it is successful in the four-dimensional case, but in higher-dimensional cases obstacles still have to be overcome. These obstacles, however, are not due to our general approach, but they lie in the fact that, at the moment, too little is known about the structure of extremal trajectories. We shall return to this question at the end of the paper. In the paper we shall give a precise description of the small-time reachable set in dimension four assuming that the constant controls $u=+1$ and $u=-1$ are not singular on the boundary of the reachable set. It can easily be seen (cf. § 4) that this is equivalent to an independence assumption on the vector fields f, g, $[f, g]$, and $[f+g, [f, g]]$, respectively, $[f-g, [f, g]]$. As a corollary we are able to improve on recent results of Bressan [4], Schättler [17], and Sussmann [21] on time-optimal control in dimension three.

Throughout this paper we will use nilpotent systems as a guide to the general situation. A system is nilpotent of order k if all brackets of orders greater than k vanish and if k is the smallest integer with this property. In a certain sense these systems play the same role as the polynomials do within the class of smooth functions. Nilpotent systems are the low-order part of the coordinate free Taylor series expansion of a general system.

To be more precise, we must define the Lie jet of system (1). At a point p the Lie jet consists of a list of the values at p of the Lie brackets of f and g written down in some prescribed order. Of course, because of the skew-symmetry and Jacobi relation

$$[f, g]+[g, f]=0, \qquad [f, [g, h]]+[g, [h, f]]+[h, [f, g]]=0,$$

we need only consider a list of distinct brackets. These brackets can be partially ordered by the total number of vector fields involved; for example, f is a bracket of order one and $[f, g]$ is of order two. The Lie jet of order k is a list of values at p of the distinct brackets of f and g of order less than or equal to k. The Lie jets of orders one through four are given below:

Order one: $\{f(p), g(p)\}$,
Order two: $\{f(p), g(p), [f, g](p)\}$,
Order three: $\{f(p), g(p), [f, g](p), [f, [f, g]](p), [g, [f, g]](p)\}$,
Order four: $\{f(p), g(p), [f, g](p), [f, [f, g]](p), [g, [f, g]](p),$
$[f, [f, [f, g]]](p), [f, [g, [f, p]]](p), [g, [g, [f, g]]](p)\}$.

If $N(k)$ is the number of distinct brackets of f and g of order k or less, then the kth-order Lie jet of (1) at p is a point in the vector bundle consisting of the Whitney sum of $N(k)$ copies of the tangent bundle.

A basic result of Krener [12], later proved in other contexts by Rothschild and Stein [15], Hermes [10], Crouch [8], Bressan [3], and Sussmann [20], [21] is that for analytic systems of the form (1), the kth-order Lie jet at p determines the trajectories emanating from p up to order $O(t^{k+1})$ and up to diffeomorphisms of the state space.

Sussmann [22], [23], Bressan [4], and Schättler [16], [17] have shown that the local structure of time-optimal controls in dimension two or three is determined in nondegenerate situations by the second, respectively, third-order Lie jet at a reference point. In degenerate situations higher-order jets need to be considered [16], [17], [23].

On the basis of these results we might conjecture that in nondegenerate situations the kth-order Lie jet at p determines the structure of the set of small-time reachable points where the Hörmander or controllability condition is satisfied, i.e., the rank of the kth-order Lie jet at p equals the dimension of the state space. And maybe the qualitative structure of the reachable set can be obtained by looking at a kth-order nilpotent approximation. Unfortunately, as we mention in the last section, these conjectures are not completely true, but they do motivate much of our work.

The paper is organized as follows. The next section reviews the Pontryagin Maximum Principle as applied to the system (1). This also gives us a chance to introduce some notation and terminology. In § 3, we will describe the main ideas and outline the general structure of our techniques by looking at the trivial two-dimensional case. We will also give a brief proof of Lobry's three-dimensional result. The main part of the paper is § 4, where we determine the geometric structure of the small-time reachable set for the nondegenerate four-dimensional system (assuming that both quadruples $(f, g, [f, g], [f+g, [f, g]])$ and $(f, g, [f, g], [f-g, [f, g]])$ consist of independent vectors at p). We also draw the obvious corollaries about time-optimal control in dimension three. Section 5 concludes with a brief discussion of the free nilpotent five-dimensional system and explains why the general nondegenerate five-dimensional case is different from this one.

2. The maximum principle. The Maximum Principle [13] gives necessary conditions for a point to lie on the boundary of the reachable set. Let $u(\cdot)$ be an admissible control defined on an interval $[0, T]$ and let $x(\cdot)$ be the corresponding trajectory starting at p. If $x(T) \in \partial$ Reach (p), then $x(t) \in \partial$ Reach (p) for all $t \in [0, T]$ and there exists an absolutely continuous curve $\lambda : [0, T] \to \mathbb{R}^n$, which does not vanish anywhere such that

$$\dot{\lambda}(t)^T = -\lambda(t)^T (Df(x(t)) + Dg(x(t)) \cdot u(t)), \tag{2}$$

$$\langle \lambda(t), g(x(t)) \rangle u(t) = \operatorname*{Min}_{|v| \leqq 1} \langle \lambda(t), g(x(t)) \rangle v, \tag{3}$$

$$H = \langle \lambda(t), f(x(t)) + g(x(t)) u(t) \rangle \equiv 0 \tag{4}$$

almost everywhere on $[0, T]$. (We write vectors as columns, $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner product on $\mathbb{R}^n$, and Df and Dg denote the Jacobian matrices of f and g, respectively.) Any trajectory for which an adjoint variable $\lambda(\cdot)$ exists such that (2)-(4) are satisfied is called an extremal trajectory. The optimality condition (3) determines the control $u(t)$ whenever $\phi(t) := \langle \lambda(t), g(x(t)) \rangle \neq 0$; ϕ is called the switching function and $u \equiv -1$ $(u \equiv +1)$ on intervals where ϕ is positive (negative). Trajectories corresponding to these constant controls are called bang arcs and are denoted by X $(=f-g)$ and Y $(=f+g)$, respectively. A concatenation of bang arcs is a *bang-bang trajectory*. Observe that $\langle \lambda(t), f(x(t)) \rangle = 0$ at switching times t, i.e., where $\langle \lambda(t), g(x(t)) \rangle = 0$. At these times (3) gives no information about the optimal control. If, however, ϕ vanishes on an open interval I, then all the derivatives of ϕ also vanish on I and this may determine the control u. We have

$$\dot{\phi}(t) = \langle \lambda(t), [f, g](x(t)) \rangle,$$

$$\ddot{\phi}(t) = \langle \lambda(t), [f+gu, [f, g]](x(t)) \rangle,$$

and if $\langle \lambda(t), [g, [f, g]](x(t)) \rangle$ does not vanish on I, we can solve for u in $\ddot{\phi} = 0$ as follows:

$$u(t) = -\frac{\langle \lambda(t), [f, [f, g]](x(t)) \rangle}{\langle \lambda(t), [g, [f, g]](x(t)) \rangle}.$$

A control of this type is called singular and the corresponding trajectory is a *singular arc.*

This suggests that concatenations of bang and singular arcs are the natural candidates for trajectories in the boundary of the reachable set (but of course no such regularity statement can be drawn from the Maximum Principle alone). We denote concatenations of bang and singular arcs by the corresponding letter sequence; for instance, we simply write XSY for a concatenation of an X-arc, followed by a singular arc and a Y-trajectory, etc.

3. The main ideas of the technique: the nondegenerate two- and three-dimensional cases. In this section we analyze the (well-known) structure of small-time reachable sets in a nondegenerate situation in dimensions two and three. These cases are easy and give us an opportunity to outline the general ideas of our technique without getting preoccupied with technical details.

Suppose Σ is a system of the form (1) in dimension two and assume that f and g are independent at a reference point p (see Fig. 1). It is clear how the small-time reachable set from p will look. If we let Γ^+ (respectively, Γ^-) be the integral curves of the vector fields $f+g$ (respectively, $f-g$) for positive times, then for sufficiently small T, Reach $(p, \leqq T)$ is the union of Γ^+, Γ^-, and the open sector R between Γ^+ and Γ^- into which $f(p)$ points. It is easy to see that any point in R is reachable from p; for instance, if $q \in R$, just run a trajectory of Σ corresponding to the control $u \equiv +1$ backward in time until it hits Γ^-. The important point is that this is all of the small-time reachable set. This follows immediately from the Maximum Principle since only trajectories corresponding to the constant controls $u \equiv +1$ or $u \equiv -1$ can lie in the boundary of the reachable set. (There cannot be a junction, since then both $\langle \lambda(t), f(x(t)) \rangle$ and $\langle \lambda(t), g(x(t)) \rangle$ vanish, contradicting the nontriviality of λ.)

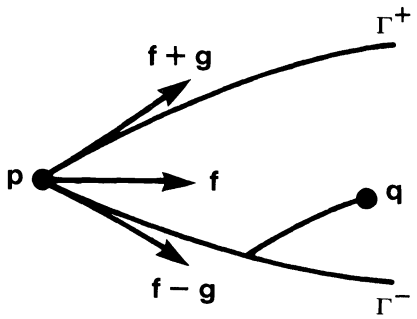

FIG. 1

Generalized to higher dimensions, the quintessence of this argument is to have two hypersurfaces Γ^* and Γ_* which are generated by extremal trajectories, have a common relative boundary and "enclose" a region R. Then, to prove that R is actually the reachable set Reach $(p, \leqq T)$, we must show (i) trajectories cannot leave R through Γ^* or Γ_*, and (ii) all points in the sector are reachable. The latter is immediate if we have a drift vector field f with $f(p) \neq 0$. This is exactly the same argument as in the

two-dimensional case. Take any point q inside R and run a trajectory of Σ corresponding to the control $u \equiv 0$ (or for that matter corresponding to any control) backward in time. Since $f(p) \neq 0$, this trajectory will hit Γ^* or Γ_*. So basically (i) must be checked; this is mostly a matter of computing tangent spaces, as will be shown below. This is the general strategy of our technique.

All technical issues left aside for a moment, the key question is how to come up with the surfaces Γ^* and Γ_*. We propose an inductive procedure. Let us explain it at the next step, which is the case of a three-dimensional system Σ, where we assume that f, g, and $[f, g]$ are independent at a reference point p. (This is the example considered by Lobry [14].)

Choose coordinates $x = (x_1, x_2, x_3)$ such that $\langle dx, (f(p), g(p), [f, g](p)\rangle = \mathrm{Id}$, the identity matrix. The projection of Σ into the (x_1, x_2)-plane is then the two-dimensional system considered above and we know the structure of its small-time reachable set. Our aim is to find two hypersurfaces Γ^* and Γ_* consisting of extremal trajectories that project onto the reachable set $\tilde{R}$ of the two-dimensional system in dimension three. If Γ^* and Γ_* have a common relative boundary that projects onto $\partial\tilde{R}$ and if Γ^* and Γ_* do not intersect in their relative interior, then it is clear that these surfaces "enclose" a region R. Then we must check whether trajectories can leave R. If this is impossible, R is the small-time reachable set.

The Maximum Principle gives preliminary information about Γ^* and Γ_* because it describes necessary conditions for trajectories to lie in the boundary of the reachable set. In this three-dimensional case it actually determines Γ^* and Γ_* precisely, but in higher dimensions this is no longer true. It is then that we will use nilpotent systems as our guide to find candidates for Γ^* and Γ_*. More on that appears in § 4.

Now that we have outlined the general approach, let us also illustrate the basic technical arguments by reproving Lobry's result. It follows from the Maximum Principle that all trajectories that lie on the boundary of the reachable set are bang-bang. For, if the switching function vanishes at some t, i.e., if $\langle \lambda(t), g(x(t))\rangle = 0$, then also $\langle \lambda(t), f(x(t))\rangle = 0$, and hence $\dot{\phi}_\Gamma(t) = \langle \lambda(t), [f, g](x(t))\rangle$ cannot vanish by the independence of f, g, and $[f, g]$ and the nontriviality of λ. For dimensionality reasons it is therefore reasonable to consider the following two surfaces as candidates for Γ^* and Γ_*:

$$\Gamma^* = \{p \exp(s_1(f-g)) \exp(s_2(f+g)) \colon s_i \geqq 0,\ s_1 + s_2 \text{ small}\},$$

$$\Gamma_* = \{p \exp(t_1(f+g)) \exp(t_2(f-g)) \colon t_i \geqq 0,\ t_1 + t_2 \text{ small}\}.$$

We write flows of vector fields as exponentials and we let the diffeomorphisms act on the right, i.e., $p \exp(tf)$ denotes the point obtained by following the integral curve of f that passes through p at time zero for t units of time.

It is clear that Γ^* and Γ_* are two-dimensional surfaces with boundary. In both cases the boundary consists of the two curves corresponding to the trajectories of $f+g$ and $f-g$ and the point p. Furthermore, by the Campbell–Hausdorff formula [11]

$$\begin{aligned} &p \exp(s_1(f-g)) \exp(s_2(f+g)) \\ &\quad = p \exp((s_1+s_2)f + (s_2-s_1)g + s_1 s_2[f, g] + s_1 s_2 \cdot O(T)), \end{aligned}$$

$$p \exp((t_1(f+g)) \exp(t_2(f-g)) = p \exp(t_1+t_2)f + (t_1-t_2)g - t_1 t_2[f, g] + t_1 t_2 \cdot O(T))$$

where $O(T)$ stands for terms that are linear in the total time T. This shows that Γ^* and Γ_* do not intersect in their relative interior. So Γ^* and Γ_* enclose a region R.

To prove that the enclosed sector R is the small-time reachable set we must show that there cannot be any other points in the reachable set. As in the two-dimensional case we have two options: either we show that we have exhausted all trajectories that

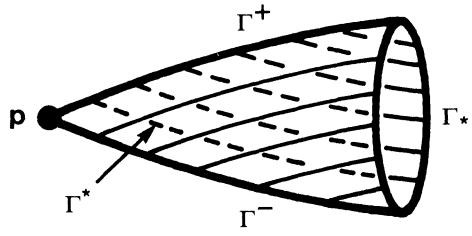

FIG. 2

possibly can lie on the boundary of the reachable set, or we show that trajectories starting at points on Γ^*, Γ_*, Γ^+, or Γ_- cannot leave $R \cup \Gamma^* \cup \Gamma_* \cup \Gamma^+ \cup \Gamma_-$. As it turns out, this is the same argument, only viewed differently.

Let us first show that we have exhausted all possible trajectories that can lie in the boundary of the small-time reachable set, i.e., that such a trajectory is bang-bang with at most one switching. Let γ be a bang-bang trajectory with two switches, say of the form XYX, with junctions p_0 and p_1 at times $t_0 < t_1$. If $\bar{\lambda} = \lambda(t_1)$, then we have $\langle \bar{\lambda}, g(p_1) \rangle = 0$ and $\langle \bar{\lambda}, f(p_1) \rangle = 0$. Also $\langle \lambda(t_0), g(p_0) \rangle = 0$ or, equivalently, if we move g ahead along the flow of the vector field Y we get $\langle \bar{\lambda}, \exp(-(t_1 - t_0) \operatorname{ad} Y) X(p_0) \rangle = 0$. But $\bar{\lambda} \neq 0$ and so these three vectors are dependent: p_0 and p_1 are *conjugate points* (Sussmann [22]). Therefore

$$X(p_1) \wedge Y(p_1) \wedge \exp(-\Delta t \operatorname{ad} Y) X(p_0) = 0$$

i.e., $X(p_1) \wedge Y(p_1) \wedge [X, Y](p_1) + O(\Delta t) = 0$, where $\Delta t = t_1 - t_0$. But such a relation cannot hold in small time by the independence of X, Y, and $[X, Y]$. Similarly it follows that YXY-concatenations cannot satisfy the Maximum Principle.

This computation can also be viewed in the following way. Define a map $F: (t_1, t_2, t_3) \mapsto p \exp(t_1 X) \exp(t_2 Y) \exp(t_3 X)$ for t_i small. Then this map has full rank if $t_i > 0$. For, if we compute the tangent space to the image, but pull back to $p \exp(t_1 X) \exp(t_2 Y)$, we get exactly the vectors $\exp(-t_2 \operatorname{ad} Y) X$, Y, and X. Therefore $F(t_1, t_2, t_3)$ is an interior point of the reachable set. Finally, if we pull back the tangent space one step further to $p \exp(t_1 X)$ we have the vectors X, Y, and $\exp(t_2 \operatorname{ad} Y) X = X - t_2[X, Y] + O(t_2^2)$. The minus sign at $[X, Y]$ implies that X-trajectories point inside R at points on Γ^*. Similarly, it follows that Y-trajectories steer the system into R from Γ^*. And this proves that trajectories of the system cannot leave R through Γ^*, Γ_*, Γ^+, or Γ_-. (Because of the Maximum Principle we can restrict ourselves to just looking at these regular controls instead of having to consider arbitrary measurable functions. For, if any trajectory would leave R, then there will also have to be additional trajectories lying on the boundary of the reachable set and these must be bang-bang.)

The structure of the small-time reachable set as a stratified set can easily be described using the following notation. For $n \in \mathbb{N}$ let

$$S_{n,-} := \{ p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \cdots \exp(s_n B) : s_i > 0,\ B = X \text{ if } n \text{ is odd},\ B = Y \text{ if } n \text{ is even} \},$$

$$S_{n,+} := \{ p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \cdots \exp(t_n B) : t_i > 0,\ B = X \text{ of } n \text{ is even } B = Y \text{ if } n \text{ is odd} \}.$$

In a nondegenerate situation each of the $S_{n,\pm}$ is a n-dimensional smooth manifold. (Certainly this will be true in all the cases we consider here.) In the three-dimensional case the boundary of the small-time reachable set consists of the two two-dimensional

strata $S_{2,\pm}$ which have in their boundary the two one-dimensional strata $S_{1,\pm}$ and the zero-dimensional stratum $S_0 = \{p\}$. S_0 also lies in the boundary of $S_{1,\pm}$. If we restrict the total time to be $\leqq T$ we must make the obvious adjustments. In particular, we must add the strata $\hat{S}_{n,\pm} := S_{n,\pm} \cap \text{Reach}(p, T)$ for $n = 1, 2$.

4. The nondegenerate four-dimensional systems. In this section we determine the geometric structure of the small-time reachable sets from a point p for a system Σ of the form (1) in dimension four, where we assume that the constant controls $u \equiv +1$ and $u \equiv -1$ are not singular. These conditions can easily be expressed in terms of independence assumptions on f, g, and lower-order brackets of f and g. For, a constant control $u \equiv u^0$ is singular on an interval I if and only if there exists an adjoint multiplier λ such that $\langle \lambda, f \rangle$, $\langle \lambda, g \rangle$, $\langle \lambda, [f, g] \rangle$, and $\langle \lambda [f + gu^0, [f, g]] \rangle$ vanish identically on I. By the nontriviality of λ this is impossible if f, g, $[f, g]$, and $[f + gu^0, [f, g]]$ are independent. Therefore in terms of the vector fields X and Y our conditions are equivalent to

(A) X, Y, $[X, Y]$ and $[X, [X, Y]]$ are independent near p;

(B) X, Y, $[X, Y]$ and $[Y, [X, Y]]$ are independent near p.

If we write $[X, [X, Y]]$ as a linear combination of X, Y, $[X, Y]$ and $[Y, [X, Y]]$ as

$$[X, [X, Y]] = \alpha X + \beta Y + \gamma [X, Y] + \delta [Y, [X, Y]],$$

then (A) is equivalent to $\delta \neq 0$.

The cases $\delta > 0$ and $\delta < 0$ are significantly different: if $\delta > 0$ only bang-bang trajectories can lie in the boundary of the reachable set, if $\delta < 0$ singular arcs are possible. Intuitively this is clear. If u is singular on an interval I, then (omitting the arguments t and $x(t)$)

$$\begin{aligned} \ddot{\phi} &= \langle \lambda, [f + gu, [f, g]] \rangle \\ &= \tfrac{1}{4} \langle \lambda, (1-u)[X, [X, Y]] + (1+u)[Y, [X, Y]] \rangle \\ &= \tfrac{1}{4} ((1-u)\delta + (1+u)) \cdot \langle \lambda, [Y [X, Y]] \rangle \neq 0 \end{aligned}$$

and so $u = (\delta + 1)/(\delta - 1)$. This is an admissible control only if $\delta \leqq 0$. Note that the singular vector field is given in feedback form as

$$S = f + \frac{\delta + 1}{\delta - 1} g = \frac{1}{1 - \delta} X + \frac{-\delta}{1 - \delta} Y, \quad \delta < 0.$$

4.1. The totally bang-bang case: $\delta > 0$. This is the generalization of Lobry's example to dimension four. We treat only the general case here, but we remark that the structure of the small-time reachable set is the same as for a nilpotent system where f, g, $[f, g]$, and $[f, [f, g]]$ form a basis and all other brackets vanish. In appropriate coordinates the latter system is linear.

The key observation again is that the Maximum Principle precisely determines the possible trajectories that can lie in the boundary of the small-time reachable set.

LEMMA 1. *If γ is a trajectory that lies in the boundary of the small-time reachable set, then γ is bang-bang with at most two switches.*

Proof. We first exclude bang-bang trajectories with more switches. Let γ be a $YXYX$-trajectory with switching points p_1, p_2, and p_3 and let s_1, s_2, s_3, s_4 be the length of the times along the respective X-arcs or Y-arcs. At every junction we have $\langle \lambda, X(p_i) \rangle = 0$ and $\langle \lambda, Y(p_i) \rangle = 0$. This gives rise to four conditions on λ.

If $\tilde{\lambda}$ is the value of the adjoint vector at the switching time at p_2, we have

$$\langle \tilde{\lambda}, X(p_2)\rangle = \langle \tilde{\lambda}, Y(p_2)\rangle = 0,$$

$$\langle \tilde{\lambda}, \exp(-s_2 \operatorname{ad} Y)X(p_1)\rangle = 0,$$

and

$$\langle \tilde{\lambda}, \exp(s_3 \operatorname{ad} X)\, Y(p_3)\rangle = 0.$$

Again, the nontriviality of $\tilde{\lambda}$ implies that these four vectors are dependent ("conjugate points"). So we get (dividing out s_2 and s_3)

$$0 = X \wedge Y \wedge \left(\frac{\exp(s_3 \operatorname{ad} X) - 1}{s_3}\right) Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2}\right) X$$

$$(5) \qquad = X \wedge Y \wedge [X, Y] + \frac{1}{2} s_3 [X, [X, Y]] + O(s_3^2) \wedge -[X, Y] + \frac{1}{2} s_2 [Y, [X, Y]] + O(T^2)$$

$$= \frac{1}{2}\sigma(s_2, s_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])_{|p_2}$$

where T is the total time along γ and $O(T^2)$ stands for terms that are quadratic in T; σ is a smooth function of s_2 and s_3. If we express $[X, [X, Y]]$ in terms of X, Y, $[X, Y]$, and $[Y, [X, Y]]$, we see that

$$(6) \qquad \sigma(s_2, s_3) = s_2 + s_3\delta + O(T^2)$$

where δ is evaluated at p_2. In a sufficiently small neighborhood of p, δ is bounded away from zero and so the linear terms dominate quadratic remainders in small time. Hence $\sigma(s_2, s_3)$ is positive for s_i small; in particular, it cannot vanish, a contradiction.

Analogously, if $\tilde{\gamma}$ is a $XYXY$-concatenation with switching points q_1, q_2, and q_3 and if t_1, t_2, t_3, t_4 are the times along the respective trajectories, then we get

$$0 = X \wedge Y \wedge \left(\frac{\exp(-t_2 \operatorname{ad} X) - 1}{-t_2}\right) Y \wedge \left(\frac{\exp(t_3 \operatorname{ad} Y) - 1}{t_3}\right) X$$

$$(7) \qquad = \frac{1}{2}\tau(t_2, t_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])_{|q_2}$$

where

$$(8) \qquad \tau(t_2, t_3) = -t_3 - t_2\delta + O(T^2)$$

is a smooth function of t_2 and t_3 near the origin. Again, since δ is bounded away from zero near p this function is negative for small times, a contradiction.

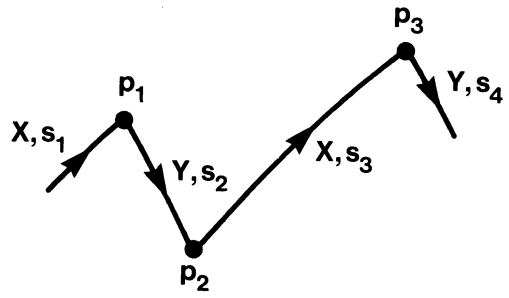

FIG. 3

It now follows that, in fact, any trajectory that lies in the boundary of the small-time reachable set is bang-bang. This is an easy but slightly technical argument. We will do it here rigorously since we will need the computations later on anyway. The point is that we do not have a priori knowledge about regularity properties of the controls, e.g., that they are piecewise constant. This is the case if and only if the zero set $Z(\phi)$ of the switching function ϕ is finite. If it were infinite, then the set N_ϕ of limit points of $Z(\phi)$ would be nonempty. In fact, it is a closed, nowhere dense, perfect set. (If $t_1 < t_2$ are points in $N(\phi)$ then, since ϕ cannot vanish identically, ϕ is different from zero somewhere in (t_1, t_2) and by continuity it is different from zero on a whole interval. It is perfect, i.e., every point $t \in N(\phi)$ is a limit point of points $t_n \in N(\phi)$, $t_n \neq t$, since $N(\phi)$ cannot have isolated points. We can see that this is so, since we know already that bang-bang trajectories with more than three switchings do not lie in the boundary of the small-time reachable set!) Suppose $t_1 < t_2$ are times in $N(\phi)$. There exists a $\tilde{t} \in (t_1, t_2)$ such that $\phi(\tilde{t}) \neq 0$. Let $\tilde{t}_1 := \sup([t_1, \tilde{t}] \cap N(\phi))$ and let $\tilde{t}_2 := \inf([\tilde{t}_1, t_2] \cap N(\phi))$. Then $\tilde{t}_1 < \tilde{t}_2$, $\tilde{t}_i \in N(\phi)$, and $Z(\phi) \cap [\tilde{t}_1, \tilde{t}_2]$ is finite. This implies that γ contains subarcs of the form $*B\cdot$ and $\cdot B*$, where B denotes a bang arc (X or Y), $\cdot$ stands for any switching, and $*$ stands for a junction in $N(\phi)$. Observe that $\dot{\phi}(t) = 0$ if $t \in N(\phi)$. We will now show that none of these concatenations can lie in the boundary of the reachable set and this will prove the lemma.

Without loss of generality we consider a concatenation of the form $*X\cdot$ with switching points p_0 and p_1 and let t be the time along X. Then, if $\tilde{\lambda}$ is the value of the adjoint vector at the switching time corresponding to p_0, we have

$$\langle \tilde{\lambda}, X(p_0) \rangle = \langle \tilde{\lambda}, Y(p_0) \rangle = \langle \tilde{\lambda}, [X, Y](p_0) \rangle = 0.$$

Also $\langle \tilde{\lambda}, \exp(-t \operatorname{ad} X) Y(p_1) \rangle = 0$ and so by nontriviality of $\tilde{\lambda}$ we again get

$$\begin{aligned} 0 &= X \wedge Y \wedge [X, Y] \wedge Y - t[X, Y] + \tfrac{1}{2}t^2[X, [X, Y]] + O(t^3) \\ &= \tfrac{1}{2}t^2(1 + O(t))(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])|_{p_0}. \end{aligned} \tag{9}$$

This cannot hold in small time. Analogously it follows that no $*B\cdot$ or $\cdot B*$ concatenation can lie in the boundary of the small-time reachable set if $\delta \neq 0$. This proves the lemma (and note that the argument is valid in general under assumptions (A) and (B)). □

It is now clear that the surfaces Γ^* and Γ_* must be as follows:

$$\Gamma^* = \{p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) : s_i \geqq 0, \text{small}\},$$

$$\Gamma_* = \{p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) : t_i \geqq 0, \text{small}\}.$$

Γ^* and Γ_* are three-dimensional surfaces with common boundary C that has precisely the structure of the boundary of the small-time reachable set in dimension three. It is the union of two two-dimensional surfaces made out of XY- and YX-trajectories respectively, glued together along the X- and Y-trajectories.

We will now show that Γ^* and Γ_* do not intersect away from C, in particular that they enclose an open region that will be the interior of the small-time reachable set.

DEFINITION. We say a point q is an entry point (respectively, an exit point) of a (closed) set S for a vector field Z if for some $\varepsilon > 0$, $S \cap \{q \exp(tZ) : -\varepsilon \leqq t \leqq 0\} = \{q\}$ (respectively, if $S \cap \{q \exp(tZ) : 0 \leqq t \leqq \varepsilon\} = \{q\}$).

LEMMA 2. *For sufficiently small T the points in Γ^* are entry points for the small-time reachable set from p for $[Y, [X, Y]]$. The points in Γ_* are exit points.*

Proof. If q is an exit (entry) point for $\operatorname{Reach}(p, \leqq T)$ that does not lie in $\operatorname{Reach}(p, T)$, i.e., exit or entry is not due to the time restriction, then the corresponding

trajectory is extremal and the adjoint multiplier satisfies the transversality condition $\langle \lambda, [Y,[X,Y]](q)\rangle \leqq 0$ ($\langle \lambda, [Y,[X,Y]](q)\rangle \geqq 0$). We claim that necessarily

$$q \in \Gamma_* \ (q \in \Gamma^*).$$

Recall that the second derivative of the switching function is given by

$$\begin{aligned} \ddot{\phi}(t) &= \langle \lambda(t), [f+gu,[f,g]](x(t))\rangle \\ &= \tfrac{1}{4}(1-u(t))\langle \lambda, [X,[X,Y]](x(t))\rangle \\ &\quad + \tfrac{1}{4}(1+u(t))\langle \lambda, [Y,[X,Y]](x(t))\rangle. \end{aligned} \tag{10}$$

Expressing $[X,[X,Y]]$ in terms of X, Y, $[X,Y]$, and $[Y,[X,Y]]$, we get a linear combination of terms $\langle \lambda, X\rangle$, $\langle \lambda, Y\rangle$, $\langle \lambda, [X,Y]\rangle$, and $\langle \lambda, [Y,[X,Y]]\rangle$, where the coefficient at $\langle \lambda, [Y,[X,Y]]\rangle$ is

$$\tfrac{1}{2}(1-u)\delta + \tfrac{1}{2}(1+u) \geqq \operatorname{Min}(1,\delta) > 0.$$

Suppose γ is a bang-bang trajectory with two junctions. Then the two junctions determine a multiplier λ up to a positive constant multiple. Normalize such that $\|\lambda(0)\|_2 = 1$. Because γ has two junctions $\langle \lambda, X\rangle$, $\langle \lambda, Y\rangle$, and $\langle \lambda, [X,Y]\rangle$ vanish somewhere on $[0,T]$, $T = t_1+t_2+t_3$. For sufficiently small T these functions will be bounded in absolute value on $[0,T]$ by any $\varepsilon > 0$. Because of (B) $|\langle \lambda(t), [Y,[X,Y]](x(t))\rangle|$ can be bounded away from zero on $[0,T]$. By choosing ε, i.e., T small enough, $\langle \lambda, [Y,[X,Y]]\rangle$ dominates all other terms in (10), that is, we have in small time: $\ddot{\phi}$ has constant sign equal to sign ($\langle \lambda, [Y,[X,Y]]\rangle$). But $\langle \lambda, [Y,[X,Y]]\rangle > 0$ allows only for XYX-trajectories and $\langle \lambda, [Y,[X,Y]]\rangle < 0$ permits only YXY-concatenations. This proves our claim.

We still need to show that points in Γ^* and Γ_* in fact have these optimization properties. Suppose γ is a XYX trajectory. Then the tangent space at the endpoint is spanned by X, $\exp(-t_3 \operatorname{ad} X)Y$ and $\exp(-t_3 \operatorname{ad} X)\exp(-t_2 \operatorname{ad} Y)X$. Note that $[Y,[X,Y]]$ always points to one side of the tangent space since

$$\begin{aligned} &X \wedge \exp(-t_3 \operatorname{ad} X)Y \wedge \exp(-t_3 \operatorname{ad} X)\exp(-t_2 \operatorname{ad} Y)X \wedge [Y,[X,Y]] \\ &= -t_2\left(X \wedge \exp(-t_3 \operatorname{ad} X)Y \wedge \exp(-t_3 \operatorname{ad} X)\left(\frac{\exp(-t_2 \operatorname{ad} Y)-1}{-t_2}\right)X \right. \\ &\qquad \left. \wedge [Y,[X,Y]]\right) \\ &= t_2(X \wedge Y - t_3[X,Y] + O(t_3^2) \wedge [X,Y] + O(T) \wedge [Y,[X,Y]]) \\ &= t_2(1+O(T))(X \wedge Y \wedge [X,Y] \wedge [Y,[X,Y]]). \end{aligned} \tag{11}$$

If we write the defining equations for Γ^* and Γ_* in terms of canonical coordinates of the second kind, that is, as products of the flows of the vector fields X, Y, $[X,Y]$, $[Y,[X,Y]]$ in the form

$$p \exp(x_1 X)\exp(x_2 Y)\exp(x_3[X,Y])\exp(x_4[Y,[X,Y]]), \tag{12}$$

then this implies that we can think of Γ^* as the graph of a function $x_4 = \psi(x_1, x_2, x_3)$. It also follows from (12) that the integral curve of $[Y,[X,Y]]$ through p and the compact set Reach (p, T) are disjoint for small positive T. Therefore, given T, there exists a $\tilde{T} \leqq T$ with the following property. Any integral curve of $[Y,[X,Y]]$ that passes through a point on $\Gamma^*(\tilde{T})$, the set of all trajectories in Γ^* of total time $\leqq \tilde{T}$, does not meet Reach (p, T). This implies that the points on $\Gamma^*(\tilde{T})$ are entry points for the small-time reachable set. For, if $q \in \Gamma^*(\tilde{T})$ is not an entry point, then by compactness

there exists an entry point of Reach $(p, \leqq T)$ of the form $q \exp r[Y, [X, Y]]$. Since this flow does not meet Reach (p, T) this point must lie on Γ^* and this contradicts the graph property. Analogously the result follows for Γ_*. □

An easy computation shows that, if Γ^* and Γ_* would intersect away from C, then it would have to happen transversally. This would contradict Lemma 2.

The geometric structure of the small-time reachable set is now clear. It is the exact analogue of Figs. 1 and 2 in four dimensions. Its boundary consists of the surfaces Γ_* and Γ^* that match up along C, the set of points reachable by a bang-bang trajectory with at most one switch. The open region enclosed by Γ^* and Γ_* is the interior of the reachable set. A stratification of its boundary is given by S_0 and $S_{n,\pm}$ for $n = 1, 2, 3$ (see § 3).

Remark. This qualitative structure of the small-time reachable set for a totally bang-bang system generalizes to arbitrary dimensions under the conditions of Krener's and Sussmann's nonlinear bang-bang theorem [19]. Suppose that the vector fields f and $\mathrm{ad}^i f(g)$, $i = 0, \cdots, n-1$ are independent at p and that for $i = 0, \cdots, n-1$ there exist smooth functions α_{ij} and β_i with $|\beta_i(p)| < 1$ such that

$$[g, \mathrm{ad}^i f(g)] = \sum_{j=0}^{i} \alpha_{ij} \mathrm{ad}^j f(g) + \beta_i \mathrm{ad}^{i+1} f(g).$$

Then it follows that for sufficiently small-time T all trajectories that lie in the boundary of the reachable set from p are bang-bang with at most n switchings. A stratification of the boundary is given by the strata $S_0 = \{p\}$ and $S_{k,\pm}$, $k = 1, \cdots, n$. In particular, points in $S_{n,+}$ are exit points of the reachable set for $(-1)^{n-1} \mathrm{ad}^{n-1} f(g)$, points in $S_{n,-}$ are entry points. Given the results on the structure of trajectories in the boundary, this is a straightforward generalization of the argument above. All the difficult work has been carried out by Sussmann in [19], specifically in the proof of Lemma 3 there.

4.2. The bang-bang singular case: $\delta < 0$. This case is a nontrivial extension of Lobry's result. Here not all the extremal trajectories actually lie in the boundary of the small-time reachable set. It is therefore not clear how we should choose Γ^* and Γ_*. We now use the structure of the small-time reachable set for the corresponding free nilpotent system as a guide. The only reasonable nilpotent approximation to choose is one where all brackets of orders greater than or equal to 4 vanish. Note that f, g, $[f, g]$, and $[g, [f, g]]$ are always independent in this case. Since we want to work with a system as simple as possible, we also assume $[f, [f, g]] \equiv 0$. This is an equality relation in the third-order Lie jet, but in a slightly more general setup (weighted Lie algebra) this would be a free nilpotent system. Therefore we refer to this system as the "free" nilpotent case. We will first analyze a model of this "free" nilpotent case, and then we will show that the general case has the same qualitative behavior.

4.2.1. The reachable set in the "free" nilpotent case. To simplify some computations we restrict ourselves to the following model $\tilde{\Sigma}$:

$$\dot{x}_0 = 1, \quad \dot{x}_1 = u, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = \tfrac{1}{2} x_1^2. \tag{13}$$

Note that $[g, f](x) = (\partial/\partial x_2) + x_1(\partial/\partial x_3)$, $[g, [g, f]] \equiv \partial/\partial x_3$ and all other brackets vanish identically. It is clear that the qualitative structure of the reachable set from the origin at any time is the same as for the small-time reachable set: one is a rescaling of the other. (If u is a control defined on $[0, T]$ and x is the corresponding trajectory, then the time 1 reachable set can be obtained from the time T reachable set by letting $\bar{u}(t) := u(t/T)$ and $\bar{x}_i(t) := T^i x_i(t/T)$ for $i = 1, 2, 3$.) To determine the reachable set it therefore suffices to look at time slices $T =$ constant, and without loss of generality we can assume $T = 1$.

If $\lambda = (\lambda_0, \lambda_1, \lambda_2, \lambda_3)^T$ is an adjoint vector for an extremal trajectory $x(\cdot)$, then λ_1 is the switching function and

$$\dot{\lambda}_1 = -\lambda_2 - \lambda_3 x_1, \quad \dot{\lambda}_2 = 0, \quad \dot{\lambda}_3 = 0,$$

and in particular $\ddot{\lambda}_1 = \lambda_3 u$, i.e., $u \equiv 0$ is the only singular control. Note that, if $\lambda_3 = 0$, then λ_1 is a linear function and the extremal trajectory is uniquely determined. By a theorem of Bressan [2] this implies that the reachable set is convex in direction of $(0, 0, 0, 1)^T$ or equivalently in the direction of $[g, [g, f]] = \frac{1}{2}[X, [X, Y]]$; that is, if (p_0, p_1, p_2, a) and (p_0, p_1, p_2, b) lie in the reachable set, then the whole segment $\{(p_0, p_1, p_2, c)\colon a \leqq c \leqq b\}$ lies in the reachable set. It is therefore clear what the surfaces Γ^* and Γ_* have to be: Γ^* consists of trajectories which are exit points for $[X, [X, Y]]$ and Γ_* of those which are entry points. Equivalently, we can speak of trajectories that maximize/minimize the coordinate x_3.

For extremal trajectories that give rise to entry/exit points for $[X, [X, Y]]$, an additional transversality condition was to hold. One of the directions $\pm[X, [X, Y]]$ can be separated from an approximating cone to the reachable set at this point. In our case these conditions simply say that $\lambda_3 \geqq 0$ for trajectories that minimize x_3 and $\lambda_3 \leqq 0$ for those that maximize x_3. In particular $\lambda_3 = 0$ for those that do both and these trajectories are bang-bang with at most one switching. So again the common boundary of Γ^* and Γ_* will be a set C that has the structure of the boundary of the small-time reachable set in dimension three.

We now determine Γ_*. We can assume $\lambda_3 > 0$ and without loss of generality normalize λ_3 to 1. Thus, $\ddot{\lambda}_1 = -u$ and so λ_1 is strictly convex and positive along X, strictly concave and negative along Y. Singular controls satisfy the generalized Legendre–Clebsch condition [13]: $\langle \lambda, [g, [f, g]] \rangle = -\lambda_3 < 0$. It follows that the only extremal trajectories are concatenations of a bang arc, followed by a singular arc and another bang arc. We now restrict to the time slice $T = 1$. Define

$$\Gamma_{-0-} := \{0 \exp(s_1 X) \exp(s_2 f) \exp(s_3 X)\colon s_i \geqq 0,\ s_1 + s_2 + s_3 = 1\},$$

$$\Gamma_{-0+} := \{0 \exp(s_1 X) \exp(s_2 f) \exp(s_3 Y)\colon s_i \geqq 0,\ s_1 + s_2 + s_3 = 1\},$$

$$\Gamma_{+0-} := \{0 \exp(t_1 Y) \exp(t_2 f) \exp(t_3 X)\colon t_i \geqq 0,\ t_1 + t_2 + t_3 = 1\},$$

$$\Gamma_{+0+} := \{0 \exp(t_1 Y) \exp(t_2 f) \exp(t_3 Y)\colon t_i \geqq 0,\ t_1 + t_2 + t_3 = 1\}.$$

We will show that these are two-dimensional surfaces with boundary which match up and together form Γ_* with

$$\partial\Gamma_* = \{0 \exp(s_1 X) \exp(s_2 Y)\colon s_i \geqq 0,\ s_1 + s_2 = 1\}$$
$$\cup \{0 \exp(t_1 Y) \exp(t_2 X)\colon t_i \geqq 0,\ t_1 + t_2 = 1\}.$$

LEMMA 3. *Each of the sets $\Gamma_{\pm 0\pm}$ is a two-dimensional surface with boundary. For any two of them the images of the open simplices are disjoint. Furthermore,*

$$\Gamma_{-0-} \cap \Gamma_{-0+} = \Gamma_{-0} = \{0 \exp(s_1 X) \exp(s_2 f)\colon s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{-0-} \cap \Gamma_{+0-} = \Gamma_{0-} = \{0 \exp(s_1 f) \exp(s_2 X)\colon s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{-0-} \cap \Gamma_{+0+} = \Gamma_0 = \{0 \exp(sf)\colon 0 \leqq s \leqq 1\} = \Gamma_{-0+} \cap \Gamma_{+0-},$$

$$\Gamma_{-0+} \cap \Gamma_{+0+} = \Gamma_{0+} = \{0 \exp(s_1 f) \exp(s_2 Y)\colon s_i \geqq 0,\ s_1 + s_2 = 1\},$$

$$\Gamma_{+0-} \cap \Gamma_{+0+} = \Gamma_{+0} = \{0 \exp(s_1 Y) \exp(s_2 f)\colon s_i \geqq 0,\ s_1 + s_2 = 1\}.$$

Graphically, these relations can be illustrated as shown in Fig. 4.

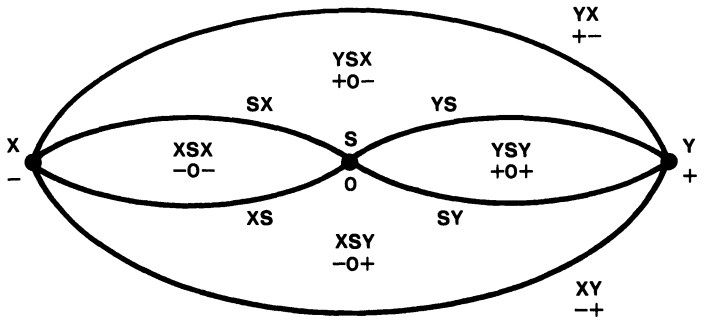

FIG. 4

The proof of the lemma consists of straightforward computations that we shall only illustrate in one case. It is easy to see that all the maps are regular with rank 2 in the interior, and it is clear how the maps behave on the boundary. So the $\Gamma_{\pm 0\pm}$ are two-dimensional surfaces with boundary. To prove that the images of the open simplex under different maps are disjoint, we choose a way that does not use the specific form of the equations, but works with a basis provided by the vector fields f, g, $[f, g]$, and $[g, [f, g]]$. This also gives an idea how the analogous argument in the general case runs. We rewrite the defining equations in terms of canonical coordinates of the second kind as products of the flows of the vector fields f, g, $[f, g]$, and $[g, [f, g]]$. Since in this case

$$\exp(f+g) = \exp([g,[f,g]]/3)\exp([f,g]/2)\exp(g)\exp(f), \tag{14}$$

we get, for instance, for Γ_{+0+}:

$$\begin{aligned}
&0\exp(t_1(f+g))\exp(t_2 f)\exp(t_3(f+g))\\
&\quad = 0\exp((\tfrac{1}{3}t_1^3[g,[f,g]])\exp(\tfrac{1}{2}t_1^2[f,g])\exp(t_1 g)\exp((t_1+t_2)f))\\
&\qquad \times\exp(\tfrac{1}{3}t_3^3[g,[f,g]])\exp(\tfrac{1}{2}t_3^2[f,g])\exp(t_3 g)\exp(t_3 f)\\
&\quad = 0\exp((\tfrac{1}{6}(t_1+t_3)^3 + t_2t_3(t_1+\tfrac{1}{2}t_3))[g,[f,g]])\exp((\tfrac{1}{2}(t_1+t_3)^2 + t_2t_3)[f,g])\\
&\qquad \times\exp((t_1+t_3)g)\exp(f).
\end{aligned}$$

Analogously we have for Γ_{-0+}:

$$\begin{aligned}
&0\exp(s_1(f-g))\exp(s_2 f)\exp(s_3(f+g))\\
&\quad = 0\exp((\tfrac{1}{3}s_1^3 - s_1^2s_3 + \tfrac{1}{3}s_3^3 + \tfrac{1}{2}s_2s_3^2 - s_1s_2s_3)[g,[f,g]])\\
&\qquad \times\exp((-\tfrac{1}{2}s_1^2 + \tfrac{1}{2}s_3^2 + (s_1+s_2)s_3)[f,g])\exp((s_3-s_1)g)\exp(f).
\end{aligned}$$

A simple computation shows that the equations we obtain by equating the coordinates have no positive solution. Similarly this is shown for all pairs of surfaces. The statements about the intersections are then clear. □

This shows that Γ_* is a two-dimensional stratified set with its one-dimensional relative boundary $\partial\Gamma_*$ made out of bang-bang trajectories with at most one switching. Figure 4 gives a precise description of the stratification. We now show that the points on Γ_* are, in fact, the points that have the smallest x_3 coordinate among all points of Reach $(0, 1)$ with a fixed (x_0, x_1, x_2).

Let us first compute the tangent spaces to the surfaces $\Gamma_{\pm 0\pm}$. Note that in each case the pullback of the tangent space to the endpoint of the singular arc simply consists of the space spanned by the vectors g and $[f, g]$ evaluated there (remember that we are working in the time slice $T=1$). This implies that $[X,[X, Y]]=2[g,[g,f]]$ always points to one side of the tangent space. In fact,

$$\exp(-t \operatorname{ad}(f \pm g))g \wedge \exp(-t \operatorname{ad}(f \pm g))[f, g] \wedge [g,[f, g]] \equiv 1(g \wedge [f, g] \wedge [g,[f, g]]).$$

In the limit this also holds for the one-dimensional strata. Therefore $[g,[g,f]]$ always points to one side of the stratified surface Γ_*. It is easy to see that, in fact, we can think of Γ_* as the graph of a piecewise defined function $x_3=\psi(x_1, x_2)$. (The projections of the images onto (x_1, x_2) intersect only along the projections of the intersections of the surfaces $\Gamma_{\pm 0\pm}$.) Since we have exhausted all possible extremal trajectories that can minimize the coordinate x_3 with Γ_*, it is now clear that given $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \Gamma_*$ any other point $(x_1, x_2, x_3) \in \operatorname{Reach}(0, 1)$ with $x_1=\bar{x}_1$ and $x_2=\bar{x}_2$ must satisfy $x_3>\bar{x}_3$. This concludes the analysis of Γ_*.

Next we will determine Γ^*. Here we can assume $\lambda_3=-1$ and so $\ddot{\lambda}_1=u$, i.e., the switching function ϕ is convex when ϕ is negative and concave when ϕ is positive. This clearly suggests bang-bang extremals. However, now the situation is significantly different from all previous cases: it will turn out that the times along bang arcs are no longer free, which in turn will mean that we cannot a priori exclude bang-bang trajectories with a large number of switchings. In general, it is a very difficult problem to eliminate extremal trajectories with a large number of switchings (cf. [4] or [16]). It turns out that in our approach we do not even have to address this issue.

Let us start by showing that the times along bang arcs can no longer vary freely. Suppose we have a concatenation of a Y-trajectory followed by an X-arc with switchings at the beginning and the end $(\cdot XY\cdot)$. Call the switching points p_0, p_1, and p_2 and let s and t be the times along X and Y, respectively. Then p_0, p_1, and p_2 are conjugate points and therefore

$$\begin{aligned} 0 &= \exp(-s \operatorname{ad} X) Y \wedge X \wedge Y \wedge \exp(t \operatorname{ad} Y) X \\ &= \left(\frac{\exp(-s \operatorname{ad} X)-1}{-s}\right) Y \wedge X \wedge Y \wedge \left(\frac{\exp(t \operatorname{ad} Y)-1}{t}\right) X \\ &= X \wedge Y \wedge [X, Y]+s[g,[f, g]] \wedge [Y, X]-t[g,[f, g]] \\ &= (s-t)(X \wedge Y \wedge [X, Y] \wedge [g,[f, g]]). \end{aligned} \tag{15}$$

Hence $s=t$ and the same is true for a $\cdot YX\cdot$-concatenation. Therefore, so as not to violate the Maximum Principle, and since we do not expect any degeneracies in the structure of the reachable set, we restrict ourselves to the following two surfaces:

$$\tilde{\Gamma}^- = \{0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \colon s_i \geqq 0,\ s_1+s_2+s_3=1,\ s_1 \leqq s_2,\ s_3 \leqq s_2\},$$

$$\tilde{\Gamma}^+ = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \colon t_i \geqq 0,\ t_1+t_2+t_3=1,\ t_1 \leqq t_2,\ t_3 \leqq t_2\}.$$

Our aim is to build Γ^* out of trajectories from $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$. However, as they are at the moment, we still have too many extremal trajectories. The surfaces $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ have a nontrivial intersection $\tilde{\gamma}$. To see this let us rewrite the defining maps in terms of canonical coordinates as follows:

$$\begin{aligned} 0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) = 0 \exp(s_1 s_2(s_2-s_1)[g,[f, g]]) \exp(s_1 s_2[X, Y]) \\ \times \exp(s_2 Y) \exp((s_1+s_3) X), \end{aligned}$$

$$0\exp(t_1Y)\exp(t_2X)\exp(t_3Y)=0\exp(t_2t_3(2t_1-t_2+t_3)[g,[f,g]])\exp(t_2t_3[X,Y])$$
$$\times\exp((t_1+t_3)Y)\exp(t_2X).$$

If we equate the coordinates, it follows easily that $s_1=t_3$, $s_2=t_2$, and $s_3=t_1$. It follows that $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ also intersect along the one-dimensional curve

$$\tilde{\gamma}=\{0\exp(sX)\exp(Y/2)\exp((\tfrac{1}{2}-s)X)\colon 0\leqq s\leqq\tfrac{1}{2}\}.$$

We need to analyze the intersection more closely. Let

$$q=0\exp(s_1X)\exp(s_2Y)\exp(s_3X)\in\tilde{\gamma}.$$

Then the tangent space to $\tilde{\Gamma}^-$ at q is spanned by (recall that $s_3=1-s_1-s_2$)

$$\exp(-s_3\operatorname{ad}X)\exp(-s_2\operatorname{ad}Y)X-X=s_2([X,Y]+(2s_3-s_2)[g,[f,g]]),$$
$$\exp(-s_3\operatorname{ad}X)Y-X=2g-s_3[X,Y]-s_3^2[g,[f,g]].$$

The point q also lies on $\tilde{\Gamma}^+$ and a tangent vector to $\tilde{\Gamma}^+$ at q is

$$t=\exp(-t_3\operatorname{ad}Y)X-Y=-2g+t_3[X,Y]-t_3^2[g,[f,g]].$$

In the intersection $t_3=s_1=:s$, $s_2=\frac{1}{2}$ and $s_3=\frac{1}{2}-s$. Thus

$$T_q\tilde{\Gamma}^-\wedge t=A(2g\wedge[X,Y]\wedge[g,[f,g]])$$

where

$$A=\begin{vmatrix}1 & s-\frac{1}{2} & -(s-\frac{1}{2})^2\\ 0 & 1 & \frac{1}{2}-2s\\ -1 & s & -s^2\end{vmatrix}=2s(s-\tfrac{1}{2})\leqq 0.$$

Hence $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect transversally except at the endpoints of $\tilde{\gamma}$ ($s=0$, $s=\frac{1}{2}$). Observe that the endpoints are characterized by the condition that the conjugate point relation $s=t$ ($=\frac{1}{2}$) holds. We need to know which surface has a larger x_3-coordinate. It follows from

$$T_q\tilde{\Gamma}^-\wedge[g,[g,f]]\equiv -2g\wedge[X,Y]\wedge[g,[f,g]]$$

that t and $[g,[g,f]]$ point to the same side of $\tilde{\Gamma}^-$ at q. Observe that $x_1=0$ for points on $\tilde{\gamma}$. Since the coefficient of t at g is negative, the points of $\tilde{\Gamma}^+$ for which $x_1<0$ have a larger x_3-coordinate than those points on $\tilde{\Gamma}^-$. Conversely for $x_1>0$ the x_3-coordinate of points on $\tilde{\Gamma}^-$ is larger. Therefore we define

$$\Gamma^-:=\{0\exp(s_1X)\exp(s_2Y)\exp(s_3X)\colon s_i\geqq 0,\ s_1+s_2+s_3=1,\ s_2\geqq\tfrac{1}{2}\},$$
$$\Gamma^+:=\{0\exp(t_1Y)\exp(t_2X)\exp(t_3Y)\colon t_i\geqq 0,\ t_1+t_2+t_3=1,\ t_2\geqq\tfrac{1}{2}\}.$$

Observe that Γ^- has the Y-trajectory in its boundary and that the X-trajectory lies in the boundary of Γ^+. Define $\Gamma^*:=\Gamma^-\cup\Gamma^+$. It follows from above that $[X,[X,Y]]=2[g,[g,f]]$ always points to one side of $\tilde{\Gamma}^-$, and similarly this holds for $\tilde{\Gamma}^+$. Since $x_1\geqq 0$ for points in Γ^-, $x_1\leqq 0$ for points in Γ^+ and $x_1=0$ exactly on the intersection, it follows that Γ^* is a piecewise defined function $x_3=\psi(x_1,x_2)$.

It is obvious that $\partial\Gamma^*$ consists of all trajectories that are bang-bang with at most one switching, i.e., $\partial\Gamma^*=\partial\Gamma_*$. Graphically, the structure is illustrated in Fig. 5.

By directional convexity it is clear that the whole set R between Γ_* and Γ^* lies in Reach $(0,1)$. We need to show that it lies nowhere else. The points of $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ that we deleted lie in the interior of R. (We deleted those points on $\tilde{\Gamma}^+$, respectively, $\tilde{\Gamma}^-$ that lie below $\tilde{\Gamma}^-$, respectively, $\tilde{\Gamma}^+$ in the direction of $[X,[X,Y]]$.) But this implies that the endpoints of bang-bang trajectories with more than two switchings lie in the interior of the reachable set. Suppose we have an extremal $XYXY$-trajectory with

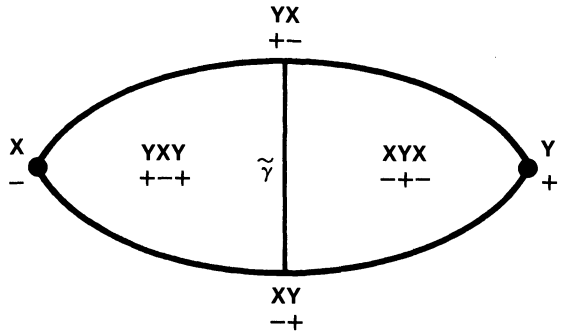

FIG. 5

times s_1, s_2, s_3, and s_4 along the trajectories. Then $s_2 = s_3$ by the conjugate point relation, and thus $s_2 < s_1 + s_3$. By the invariance of the structure of the reachable set it follows that $0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \in \text{int Reach}(0, s_1 + s_2 + s_3)$. (This is a point of the type we deleted!) Hence the trajectories that define Γ^+ and Γ^- are the only extremal trajectories that can lie on the boundary of the reachable set. This proves $R = \text{Reach}(0, 1)$.

Summary. For every time t the time $-t-$ reachable set is a stratified set that is topologically a sphere. Its boundary consists of two hemispheres $\Gamma^*(t)$ and $\Gamma_*(t)$ whose common relative boundary $\partial\Gamma^*(t)$ consists of all points reachable in time t by a bang-bang trajectory with at most one switch. $\Gamma^*(t)$ consists of all bang-bang trajectories with at most two switchings for which the time along the intermediate arc is greater than or equal to the sum of the times of the adjacent arcs. $\Gamma_*(t)$ consists of all trajectories that are concatenations of a bang arc, followed by a singular arc and another bang arc, where the times along these trajectories are free subject to $0 \leqq \text{time} \leqq t$. The stratification of its boundary is given in Figs. 4 and 5.

4.2.2. The general case. We now show that the qualitative structure of the small-time reachable set does not change in the general case. Clearly, some of the arguments will have to be adjusted; for instance, the correct generalization of the arguments using directional convexity now use the integral curves of $[X, [X, Y]]$. However, finding a general version for the explicit computations in the analysis of the bang-bang extremal trajectories is crucial.

We first define Γ_*. Recall that the singular control is given in feedback form as $u = (\delta+1)/(\delta-1)$ and since $\delta < 0$ we have no problems with u hitting the control constraint $|u| = 1$ in small time. Let $\rho = 1/(1-\delta)$, $\rho \in (0, 1)$, and let $S := f + (\delta+1)/(\delta-1)g = \rho X + (1-\rho)Y$, be the singular vector field. Define

$$\Gamma_{-s-} := \{p \exp(s_1 X) \exp(s_2 S) \exp(s_3 X) : s_i \geqq 0, \text{small}\},$$

$$\Gamma_{-s+} := \{p \exp(s_1 X) \exp(s_2 S) \exp(s_3 Y) : s_i \geqq 0, \text{small}\},$$

$$\Gamma_{+s-} := \{p \exp(t_1 Y) \exp(t_2 S) \exp(t_3 X) : t_i \geqq 0, \text{small}\},$$

$$\Gamma_{+s+} := \{p \exp(t_1 Y) \exp(t_2 S) \exp(t_3 Y) : t_i \geqq 0, \text{small}\},$$

$$\Gamma_* := \Gamma_{-s-} \cup \Gamma_{-s+} \cup \Gamma_{+s-} \cup \Gamma_{+s+}.$$

If we replace f by S in Lemma 3, then the statement stays true verbatim for $\Gamma_{\pm s\pm}$ instead of $\Gamma_{\pm 0\pm}$. (The computations are a straightforward though somewhat messy extension of the computation in the "free" nilpotent case and we omit them.) So again Γ_* is a stratified two-dimensional surface; its one-dimensional relative boundary $\partial\Gamma_*$ is made out of the bang-bang trajectories with at most one switching.

LEMMA 4. *For sufficiently small T the points on Γ_* are entry points of* Reach $(p, \leqq T)$ *for* $[X,[X,Y]]$.

Proof. The strategy is the same as in the proof of Lemma 2. We first show that the extremals on Γ_* satisfy the necessary transversality condition for entry points (which are not due to the time constraint). Then we show that Γ_* actually is a graph with the coefficient of the flow of $[X,[X,Y]]$ as dependent variable. As in Lemma 2 this suffices to prove our result.

If γ is any trajectory containing a singular arc then, for sufficiently small time, $\langle\lambda,[X,[X,Y]]\rangle$ will dominate $\langle\lambda,X\rangle$, $\langle\lambda,Y\rangle$, and $\langle\lambda,[X,Y]\rangle$, in particular, it has constant sign. Along the singular arc $\langle\lambda,[X,[X,Y]]\rangle = 2\delta/(1-\delta)\,\langle\lambda,[g,[X,Y]]\rangle$ and the generalized Legendre-Clebsch condition implies that $\langle\lambda,[X,[X,Y]]\rangle$ is positive. This shows that points in Γ_* satisfy the necessary transversality condition. An argument analogous to the one made in the proof of Lemma 1 shows that, in fact, any extremal trajectory for which $\langle\lambda,[X,[X,Y]]\rangle$ is positive has to be of the form *BSB*, that is, we have exhausted all possible candidates. To prove that indeed each point on Γ_* has the entry property, we show again that we can think of Γ_* as the graph of a piecewise defined function $x_3 = \psi(x_0, x_1, x_2)$, where (x_0, x_1, x_2, x_3) are canonical coordinates of the second kind, and x_3 is the coefficient at the flow of $[X,[X,Y]]$. Let us consider, for instance, Γ_{+s-}. It is easier to compute the pullback of the tangent space to the endpoint of the singular arc. It is spanned by X, S, and $\exp(-t_2 \operatorname{ad} S)X$. Note that $S = \rho X + (1+\rho)Y$ and it follows by induction that $\operatorname{ad}^n S(X) = \alpha_n X + \beta_n Y + \gamma_n [X,Y]$ with smooth functions $\alpha_n, \beta_n, \gamma_n$:

$$\begin{aligned}[S, \operatorname{ad}^{n-1} S(X)] &= [\rho X + (1-\rho)Y,\ \alpha_{n-1}X + \beta_{n-1}Y + \gamma_{n-1}[X,Y]] \\ &= \gamma_{n-1}\underbrace{(\rho[X,[X,Y]] + (1-\rho)[Y,[X,Y]])}_{=\rho(\alpha X+\beta Y+\gamma[X,Y])} + f, g \text{ or } [f,g] \text{ terms}.\end{aligned}$$

Also $[S,X] = [\rho X + (1-\rho)Y, X] = 2L_x(\rho)g + (\rho-1)[X,Y]$. Therefore

$$X \wedge S \wedge \exp(-t_2 \operatorname{ad} S)X = (1-\rho)^2 t_2 (1 + O(t_2)) \cdot (f \wedge g \wedge [X,Y]).$$

Now if we take the wedge-product with $[X,[X,Y]]$ pulled back along X, t_3 this yields

$$\begin{aligned}&X \wedge S \wedge \exp(-t_2 \operatorname{ad} S)X \wedge \exp(t_3 \operatorname{ad} X)([X,[X,Y]]) \\ &\quad = (1-\rho)^2 t_2 (1 + O(T)) \cdot (f \wedge g \wedge [f,g] \wedge [X,[X,Y]])\end{aligned}$$

and there are no problems with dominance since t_2 factors. Hence $[X,[X,Y]]$ always points to one side of Γ_{+s-} in the interior. Analogously it follows for the other surfaces. By continuity this also follows for the one-dimensional strata. Straightforward but slightly more tedious computations show also that the projections of the relative interiors of the sets $\Gamma_{\pm s\pm}$ onto (x_0, x_1, x_2)-space are pairwise disjoint. Therefore Γ_* is a graph in canonical coordinates. This proves the lemma. □

The analysis of the bang-bang extremals is more difficult. We start by computing the conjugate point relations. Suppose γ is a $\cdot XYX\cdot$-concatenation starting at p with junctions at p, p_1, p_2, p_3 and times s_1, s_2, s_3 along the respective trajectories. Then we have (the vector fields are evaluated at p_1):

$$\begin{aligned}0 &= X \wedge Y \wedge \left(\frac{\exp(-s_1 \operatorname{ad} X) - 1}{-s_1}\right) Y \wedge \left(\frac{\exp(s_2 \operatorname{ad} Y) - 1}{s_2}\right) X \\ &= X \wedge Y \wedge [X,Y] - \frac{1}{2}s_1 [X,[X,Y]] + O(s_1^2) \wedge -[X,Y] - \frac{1}{2}s_2 [Y,[X,Y]] + O(s_2^2) \qquad (16) \\ &= \frac{1}{2}\sigma(s_1, s_2)(X \wedge Y \wedge [X,Y] \wedge [Y,[X,Y]])|_{p_1}\end{aligned}$$

where $\sigma(s_1, s_2) = -s_1\delta - s_2 + O(2)$.

The equation $\tilde{\sigma}(s_1, s_2) = 0$ has a unique solution $\overline{s_1}(s_2)$ and in general $XYYX$-trajectories only satisfy the necessary conditions of the Maximum Principle if $s_1 \leqq \overline{s_1}(s_2)$. Note that $\varrho(0, s_2) < 0$ and so this is equivalent to $\varrho(s_1, s_2) \leqq 0$. (Using an argument analogous to (9) it can be shown that extremal trajectories do indeed have switchings at $s_1 = \overline{s_1}$, but we will not need this.) Furthermore,

$$0 = X(p_2) \wedge Y(p_2) \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2}\right) X(p_1)$$
$$\wedge \left(\frac{\exp(s_3 \operatorname{ad} X) - 1}{s_3}\right) Y(p_3)$$
$$= X \wedge Y \wedge -[X, Y] + \frac{1}{2}s_2[Y, [X, Y]] + \cdots \wedge [X, Y] + \frac{1}{2}s_3[X, [X, Y]] + \cdots$$
$$= \frac{1}{2}\tilde{\sigma}(s_2, s_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])|_{p_2}$$

where $\tilde{\sigma}(s_2, s_3) = -s_2 - s_3\delta + O(T^2)$.

Again the equation $\tilde{\sigma}(s_2, s_3) = 0$ can be solved by $\overline{s_3}(s_2)$, and YXY-concatenations only satisfy the Maximum Principle if $s_3 \leqq \overline{s_3}(s_2)$. Since $\tilde{\sigma}(s_2, 0) < 0$ this is equivalent to $\tilde{\sigma}(s_2, s_3) \leqq 0$.

Therefore we define

$$\tilde{\Gamma}^- = \{p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) : s_i \geqq 0, \text{ small}, s_2 \text{ is free},$$
$$\varrho(s_1, s_2) \leqq 0, \tilde{\sigma}(s_2, s_3) \leqq 0\}.$$

Analogously we must compute the conjugate point relations along a $\cdot YXY\cdot$-concatenation which yields

$$\tilde{\Gamma}^+ := \{p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) : t_1 \geqq 0, \text{ small } t_2 \text{ is free},$$
$$\tau(t_1, t_2) \geqq 0 \Leftrightarrow t_1 \leqq \bar{t}_1(t_2)\, \tilde{\tau}(t_2, t_3) \geqq 0 \Leftrightarrow t_3 \leqq \bar{t}_3(t_2)\}$$

where

$$\tau(t_1, t_2) = -t_1 - t_2\delta + O(T^2), \qquad \tilde{\tau}(t_2, t_3) = -t_2\delta - t_3 + O(T^2)$$

and $\bar{t}_1$ and $\bar{t}_3$ are the solutions of $\tau = 0$ and $\tilde{\tau} = 0$, respectively. $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ are three-dimensional surfaces with relative boundary made up entirely of bang-bang trajectories with at most one switch.

LEMMA 5. *The surfaces $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect along a two-dimensional surface $\hat{\Gamma}$.*

The intersection of $\hat{\Gamma}$ with the relative boundaries $\partial\tilde{\Gamma}^-$ and $\partial\tilde{\Gamma}^+$ are the following one-dimensional curves:

$$\tilde{\gamma} = \{p \exp(s_1 X) \exp(s_2 Y) : s_2 \geqq 0, \textit{small}, s_1 = \bar{s}_1(s_2)\},$$
$$\gamma = \{p \exp(t_1 Y) \exp(t_2 X) : t_2 \geqq 0, \textit{small}, t_1 = t_1^-(t_2)\}$$

(i.e., the trajectories corresponding to the conjugate points). Away from γ and $\tilde{\gamma}$ the surface entirely lies in the relative interior of $\tilde{\Gamma}^-$, respectively, $\tilde{\Gamma}^+$ and there the intersection is transversal.

Proof. We want to solve the equation

$$p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) = p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y). \tag{17}$$

Suppose a point q in the relative interior of $\tilde{\Gamma}^+$ or $\tilde{\Gamma}^-$ lies on $\hat{\Gamma}$. We claim that (16) can be solved in terms of t_1 and t_2 near q. This follows from the Implicit Function Theorem if the Jacobian with respect to (s_1, s_2, s_3, t_3) is nonsingular at q. If we compute these derivatives and pull the vectors back along X we get

$$\begin{aligned}
&\exp(-s_2 \operatorname{ad} Y)X \wedge Y \wedge X \wedge \exp(s_3 \operatorname{ad} X)Y \\
&\quad = s_2 s_3 \left(X \wedge Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2} \right) X \wedge \left(\frac{\exp(s_3 \operatorname{ad} X) - 1}{s_3} \right) Y \right) \\
&\quad = \frac{1}{2} s_2 s_3 \cdot \tilde{\sigma}(s_2, s_3)(X \wedge Y[X, Y] \wedge [Y, [X, Y]])|_{p_2 = p \exp(s_1 X) \exp(s_2 Y)}.
\end{aligned}$$

But in int $(\tilde{\Gamma}^-)$ s_2 and s_3 are positive and also $\tilde{\sigma}(s_2, s_3) < 0$ since the conjugate point relation does not hold. So we can solve in terms of t_1 and t_2. This computation shows also that $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ intersect transversally in int $(\tilde{\Gamma}^+)$ or int $(\tilde{\Gamma}^-)$.

Next we show that points q of this type exist. For that we rewrite both sides of (17) in terms of canonical coordinates of the second kind. A short computation (cf., for instance, [16]) shows that

$$\begin{aligned}
p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) &= p \exp(\tfrac{1}{2} s_1 s_2 (s_1 \delta + s_2 + O(S^2))[Y, [X, Y]]) \\
&\quad \cdot \exp(s_1 s_2 (1 + O(S))[X, Y]) \\
&\quad \cdot \exp((s_2 + O(S^3))Y) \exp((s_1 + s_3 + O(S^3))X), \\
p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) &= p \exp(\tfrac{1}{2} t_2 t_3 (2t_1 + t_3 + t_2 \delta + O(T^2))[Y, [X, Y]]) \\
&\quad \cdot \exp(t_2 t_3 (1 + O(T))[X, Y]) \\
&\quad \cdot \exp((t_1 + t_3 + O(T^3))Y) \exp((t_2 + O(T^3))X)
\end{aligned}$$

where $O(S^k)$ or $O(T^k)$ stand for terms of order greater than or equal to k in the total time, $S = s_1 + s_2 + s_3$, $T = t_1 + t_2 + t_3$, and δ is evaluated at p. Equating coefficients we get

$$\begin{aligned}
&\text{(i)} && s_1 + s_3 + O(S^3) = t_2 + O(T^3), \\
&\text{(ii)} && s_2 + O(S^3) = t_1 + t_3 + O(T^3), \\
&\text{(iii)} && s_1 s_2 (1 + O(S)) = t_2 t_3 (1 + O(T)), \\
&\text{(iv)} && s_1 s_2 (s_1 \delta + s_2 + O(S^2)) = t_2 t_3 (2t_1 + t_3 + t_2 \delta + O(T^2)).
\end{aligned} \tag{18}$$

If we assume that all switching times are comparable, i.e., of order T, then (18(i), (ii)), and

$$\text{(iv')} \quad s_1 \delta + s_2 + O(S^2) = 2t_1 + t_3 + t_2 \delta + O(T^2)$$

can easily be solved for s in terms of t modulo higher-order terms:

$$\begin{aligned}
s_1 &= t_2 + \frac{1}{\delta} t_1 + O(T^2), \\
s_2 &= t_1 + t_3 + O(T^3), \\
s_3 &= -\frac{1}{\delta} t_1 + O(T^2).
\end{aligned} \tag{19}$$

With these times the conjugate point relations cannot hold since

$$\tilde{\sigma}(s_2, s_3) = -s_2 - s_3 \delta + O(T^2) = -t_3 + O(T^2) \tag{20}$$

is negative. So the corresponding point q lies in fact in the relative interior and therefore it is possible to solve for t_3 in terms of t_1 and t_2:

$$t_3 = -t_1 - \delta t_2 + O(T^2). \tag{21}$$

This gives a solution to (18). Note that

$$t_2 = \rho T + O(T^2) = \frac{T}{1-\delta} + O(T^2). \tag{22}$$

As long as (t_1, t_2, t_3) are bounded away from the boundary of the simplex $t_1 + t_2 + t_3 \leqq T$, the times are comparable, these computations are justified, and we get a two-dimensional intersection that we can parametrize by t_1 and t_2. The problem is whether it extends all the way to the boundary. But the equations (19) and (21) are well defined for $t_1 \to 0$ (in a time-slice $t_1 + t_2 + t_3 = T$ it follows that $t_3 \to -\delta t_2 + O(t_2^2)$, i.e., to a limit of order T. By (20) this implies that the two-dimensional surface defined by these functions of (t_1, t_2) stays away from the conjugate point condition $\tilde{\sigma}(s_2, s_3) = 0$. Hence the implicit function theorem is still applicable.) Therefore $\hat{\Gamma}$ extends all the way out to $t_1 = 0$, i.e., to the XY boundary surface.

A precise characterization of $\tilde{\Gamma}^+ \cap \tilde{\Gamma}^- \cap \{p \exp(s_1 X) \exp(s_2 Y)\colon s_i \geqq 0, \text{small}\}$ is possible. Clearly these are points such that $t_1 = 0$, $t_2 = s_1$, $t_3 = s_2$, and $0 = s_3$. Since $(s_1, s_2, 0) \in \operatorname{dom} \tilde{\Gamma}^-$ we have $\varrho(s_1, s_2) \leqq 0$, and since $(0, s_1, s_2) \in \operatorname{dom} \tilde{\Gamma}^+$ we have $\tilde{\tau}(s_1, s_2) \geqq 0$. But in this case $\varrho(s_1, s_2) = \tilde{\tau}(s_1, s_2)$ (cf. (16) and the analogous formula for $\tilde{\tau}$). Therefore $\varrho(s_1, s_2) = 0$, i.e., $s_1 = \overline{s_1}(s_2)$, the conjugate point relation.

This proves that $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ extends all the way out to the XY-boundary surface and that the intersection with the XY-surface is the one-dimensional curve $\tilde{\gamma}$ consisting of the conjugate points.

Analogously we can show that (17) can also be solved in terms of s_1 and s_2 in int $(\tilde{\Gamma}^-)$. Using these formulas we can show that $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ extends all the way up to the YX-boundary surface and that the intersection of $\tilde{\Gamma}^- \cap \tilde{\Gamma}^+$ with the YX-surface consists of the curve γ. □

Note that in a time-slice $t_1 + t_2 + t_3 = T$ the qualitative geometric structure of $\tilde{\Gamma}^- \cup \tilde{\Gamma}^+$ is exactly as in the free nilpotent case. Only the condition $t_2 = T/2$ is replaced by $t_2 \doteq (1/(1-\delta))T$ (modulo higher terms) which shifts $\hat{\Gamma}$ away from the center. This is illustrated in Fig. 6.

The surface $\hat{\Gamma}$ bisects $\tilde{\Gamma}^+$ and $\tilde{\Gamma}^-$ and only one of the two components has the Y-, respectively, X-trajectory in its boundary. We define Γ^- and Γ^+ to be these components

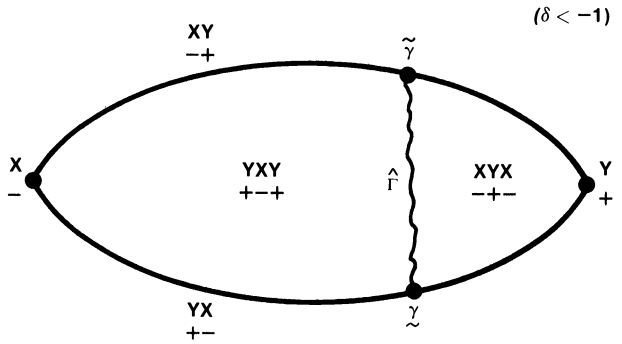

FIG. 6

and let $\Gamma^* = \Gamma^+ \cup \Gamma^-$. It is then clear that Γ^* is a three-dimensional stratified surface whose relative boundary consists of all bang-bang trajectories with at most one switching, i.e., $\partial\Gamma^* = \partial\Gamma_*$.

LEMMA 6. *The points in* Γ^* *are exit points of the small-time reachable set for* $[X, [X, Y]]$.

Proof. It is easy to see (cf. (10)) that, for sufficiently small time, all extremals on $\tilde{\Gamma}^-$ or $\tilde{\Gamma}^+$ satisfy the necessary transversality condition $\langle \lambda, [X, [X, Y]] \rangle \leqq 0$.

We show first that the points that we deleted from $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ are not exit points (see Fig. 7). Let

$$q = p \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) = p \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y)$$

be a point in the relative interior of $\hat{\Gamma}$. $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect transversally. It follows as in the proof of Lemma 2 (cf. (11)) that the XYX- and YXY-surfaces are graphs $x_4 = \psi(x_1, x_2, x_3)$ in canonical coordinates of the second kind with x_4 the coefficient at the flow of $[X, [X, Y]]$. This inherits on $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$. To prove that the parts of $\tilde{\Gamma}^-$ (respectively, $\tilde{\Gamma}^+$) that we delete are not exit points, it suffices to show that these parts lie below $\tilde{\Gamma}^+$ (respectively, $\tilde{\Gamma}^-$) in direction of $[X, [X, Y]]$.

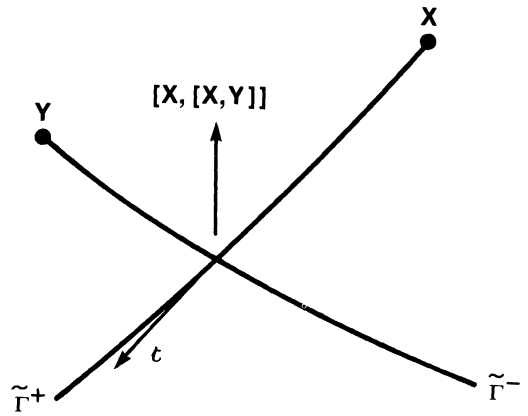

FIG. 7

The tangent space to $\tilde{\Gamma}^-$ at q is spanned by X, $\exp(-s_3 \operatorname{ad} X)Y$ and $\exp(-s_3 \operatorname{ad} X)(\exp((-s_2 \operatorname{ad} Y) - 1)/-s_2)X$. To show that the part of $\tilde{\Gamma}^+$ that we deleted lies below $\tilde{\Gamma}^-$ near q it suffices to show that $[X, [X, Y]]$ and a tangent vector t to $\tilde{\Gamma}^+$ that is oriented toward the sector of $\tilde{\Gamma}^+$ that we deleted point to opposite sides of $T_q\tilde{\Gamma}^-$. We get such a vector t if we lengthen the time along the last Y leg. (We delete the piece that contains in its boundary the trajectories corresponding to the conjugate point relation $t_3 = \bar{t}_3(t_2)$.)

Instead of computing at q we pull back all vectors along X, s_3 and get

$$\exp(+s_3 \operatorname{ad} X)(T_q\tilde{\Gamma}^-) \wedge \exp(+s_3 \operatorname{ad} X)[X, [X, Y]]$$

$$= \left(X \wedge Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y) - 1}{-s_2} \right) X \wedge \exp(s_3 \operatorname{ad} X)[X, [X, Y]] \right)$$

$$= -(\delta + O(T))(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])|_{p_2 = p \exp(s_1 X) \exp(s_2 Y)},$$

$$\exp(s_3 \operatorname{ad} X)(T_q\tilde{\Gamma}^-) \wedge \exp(s_3 \operatorname{ad} X)\, Y$$
$$= s_3\left(X \wedge Y \wedge \left(\frac{\exp(-s_2 \operatorname{ad} Y)-1}{-s_2}\right) X \wedge \left(\frac{\exp(s_3 \operatorname{ad} X)-1}{s_3}\right) Y\right)$$
$$= \frac{1}{2}\, s_3 \tilde{\sigma}(s_2, s_3)(X \wedge Y \wedge [X, Y] \wedge [Y, [X, Y]])|_{p_2}.$$

But q is a point in $\hat{\Gamma}$, and $\hat{\Gamma}$ lies entirely in the relative interior of $\tilde{\Gamma}^-$ except for the obvious boundary curves $\tilde{\gamma}$ and γ. In particular (cf. also the proof of Lemma 5) the conjugate point relation $s_3 = \bar{s}_3(s_2)$ does not hold, or equivalently, $\tilde{\sigma}(s_2, s_3) < 0$. So these wedge-products have opposite signs, which proves our claim. This also implies that the portion of $\tilde{\Gamma}^-$ that we delete lies below $\tilde{\Gamma}^+$, and since there is no other intersection this holds for all the points we deleted.

The stratified sets Γ^* and Γ_* enclose a region R that lies in the small-time reachable set. In particular, the portions of $\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ that we deleted therefore lie in the interior of the reachable set. Since these pieces contain the trajectories corresponding to the conjugate points $t_3 = \bar{t}_3(t_2)$ and $s_3 = \overline{s_3}(s_2)$, it follows that no bang-bang trajectory with more than two switchings lies in the boundary of the small-time reachable set. Hence the points in Γ^* are the only possible exit points of the small-time reachable set for $[X, [X, Y]]$. It follows from the construction of Γ^- and Γ^+ that Γ^* is also a graph. Again, the projections onto (x_1, x_2, x_3)-space are disjoint. Therefore it follows as in Lemma 2 that the points on Γ^* have the exit property for sufficiently small time. □

Finally, Γ^* and Γ_* do not intersect in their relative interiors. It is now clear how the *small-time reachable* set looks: It is the set of points enclosed by the two three-dimensional stratified surfaces Γ^* and Γ_*. Γ^* consists of bang-bang trajectories with at most two switchings such that modulo higher-order terms

$$t_1 + \delta t_2 + t_3 \leqq 0 \tag{23}$$

if t_1, t_2, and t_3 are the consecutive times along a YXY arc and

$$s_1\delta + s_2 + s_3\delta \geqq 0 \tag{24}$$

if s_1, s_2, s_3 are consecutive times along XYX. Γ_* consists of all concatenations of a bang arc, followed by a singular arc and another bang arc where the time along the trajectories is free. Γ^* and Γ_* have a common relative boundary C consisting of all trajectories that are bang-bang with at most one switching. For sufficiently small-time T a time-slice of the reachable set has exactly the same qualitative geometric structure as for the free nilpotent system (13). Furthermore, if $\delta(\cdot)$ is an integral curve of $[X, [X, Y]]$ such that $\delta(t_1)$ and $\delta(t_2)$, $t_1 < t_2$, lie in the small-time reachable set, then so does the whole curve $\delta(t)$, $t_1 \leqq t \leqq t_2$. The points on Γ_* are entry points for $[X, [X, Y]]$; the points on Γ^* are exit points.

Remark. We emphasize that the result is not what might be expected intuitively. From dimensionality we could conjecture the occurrence of bang-bang trajectories with two switchings, respectively, *BSB* trajectories in the boundary of the small-time reachable set. Also, this is essentially what was partially known from earlier results. However, we see no simple reasoning that could explain why, in fact, *some* of these *bang-bang trajectories with two switchings are not a part of the boundary.* This is only revealed by our analysis.

4.3. Time-optimal control in dimension three. Our results have immediate implications on time-optimal control in dimension three. Suppose the triples $(g, [f, g], [f+g, [f, g]])$ and $(g, [f, g], [f-g, [f, g]])$ consist of independent vectors at a point p in $\mathbb{R}^3$.

Equivalently, suppose that the constant controls $u \equiv +1$ and $u \equiv -1$ are not singular. If we augment the three-dimensional system Σ to a four-dimensional system $\hat{\Sigma}$ by introducing time as a coordinate, $\dot{x}_0 = 1$, $x_0(0) = 0$, i.e.,

$$\hat{f} = \begin{pmatrix} 1 \\ f \end{pmatrix}, \qquad \hat{g} = \begin{pmatrix} 0 \\ g \end{pmatrix},$$

then if a Σ-trajectory $x(\cdot) : [0, T] \to \mathbb{R}^3$ steering p to q is time-optimal, the augmented trajectory $\hat{x}$ lies in the boundary of the reachable set from p. The augmented system $\hat{\Sigma}$ satisfies our assumptions (A) and (B), and therefore time-optimal trajectories are bang-bang with at most two switchings or concatenations of a bang-arc, followed by a singular arc and one more bang arc. Under additional assumptions this result was obtained earlier by Bressan [4], who studied only trajectories emanating from an equilibrium point of f and by Sussmann [22] and Schättler [17] who both assumed in addition also that f, g and $[f, g]$ were independent. Our analysis shows that the vector field f is irrelevant and we do not have to make any assumptions about it. Our results are also more precise in the sense that we can exclude the optimality of those bang-bang trajectories with two switchings that violate (23) (respectively, (24)) in the bang-bang singular case. We summarize in the following corollary.

COROLLARY. *Suppose the vector fields* g, $[f, g]$ *and* $[f+g, [f, g]]$ *are independent near a reference point* $p \in \mathbb{R}^3$. *Write*

$$[f-g, [f, g]] = ag + b[f, g] + c[f+g, [f, g]]$$

and assume that c *does not vanish. Then we have in small time*:

(i) *If* $c > 0$, *then time-optimal trajectories are bang-bang with at most* 2 *switches.*

(ii) *If* $c < 0$, *then time-optimal trajectories are bang-bang with at most two switchings or are concatenations of a bang arc, a singular arc, and another bang arc. Time-optimal* XYX (*respectively*, YXY) *concatenations satisfy modulo higher-order terms*

$$c(s_1 + s_3) + s_2 \geqq 0 \qquad (\textit{resp.}, t_1 + t_3 + ct_2 \leqq 0)$$

where s_1, s_2, s_3 (*respectively*, t_1, t_2, t_3) *are the consecutive times along the bang arcs.*

5. A brief outlook to higher dimensions. We have outlined a general method to determine the structure of the small-time reachable sets and proved its effectiveness in nondegenerate cases in small dimensions. One of the difficulties that will become more and more prominent in higher dimensions is that the necessary conditions of the Maximum Principle will not restrict the class of extremal trajectories sufficiently enough to give the candidates for Γ^* and Γ_*.

Under assumptions (A) and (B) in dimension four, we could overcome this problem by taking a corresponding "free" nilpotent system of the same dimension as a guide. We do not expect this to happen in general. In fact, for the five-dimensional system Σ, where we assume that f, g, $[f, g]$, $[f, [f, g]]$, and $[g, [f, g]]$ are independent, the small-time reachable set has extremal trajectories in its boundary that do not appear in the analogous five-dimensional free nilpotent system. The reason for this lies in a qualitatively different behavior of the singular controls, specifically, in the fact that singular controls can now hit the control constraint $|u| = 1$ and may have to be terminated. Nevertheless, the free nilpotent system contains most of the information about the small-time reachable set, though it does not characterize it completely. To be more specific, we will briefly describe (without proofs) the structure of the reachable set for the free nilpotent system in dimension five and how the general case differs from it.

We take as our model:

$$\dot{x}_0 = 1, \quad \dot{x}_1 = u, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = x_2, \quad \dot{x}_4 = \tfrac{1}{2}x_1^2.$$

It is no problem whatsoever to carry out the analysis within our technique as in the construction in § 4.2.1. Now the reachable set is convex in direction of $(0, 0, 0, 0, 1)^T = [g, [g, f]]$ and Γ^*, respectively, Γ_* will consist of those trajectories that are exit, respectively, entry points.

It follows from the generalized Legendre-Clebsch condition that Γ_* contains concatenations with singular arcs, whereas Γ^* will consist of bang-bang trajectories only. Singular controls are constant, but now they can take on any value in $[-1, 1]$.

Let $\Gamma_* = \Gamma_{-u-} \cup \Gamma_{-u+} \cup \Gamma_{+u-} \cup \Gamma_{+u+}$, where

$$\Gamma_{-u-} := \{0 \exp(s_1 X) \exp(s_2(f+ug)) \exp(s_3 X) : s_i \geqq 0,\ s_1 + s_2 + s_3 = 1,\ u \in [-1, 1]\},$$

etc. (By the invariance property of the reachable set we can restrict to the time-slice $T = 1$.) The points on Γ_* are precisely the ones that minimize the coordinate x_4.

For a fixed value u_0 of the singular control, $-1 < u_0 < +1$, the qualitative structure of $\Gamma_{*,u_0} = \Gamma_*$ restricted to values $u = u_0$ is precisely as in 4.2.2, Fig. 4 (see Fig. 8).

For $u_0 = +1$, $\Gamma_{-u-} \upharpoonright u = 1$ reduces to Γ_{-+-} and all other strata become trivial whereas for $u_0 = -1$, $\Gamma_{+u+} \upharpoonright u = -1 = \Gamma_{+-+}$ and the remaining strata are trivial. For each of these two-dimensional surfaces (u_0 fixed) the relative boundary consists of all bang-bang trajectories with at most one switching. The surfaces Γ_{*,u_0} themselves interpolate between Γ_{+-+} for $u_0 = -1$ and Γ_{-+-} for $u_0 = 1$. Topologically Γ_* is a stratified sphere

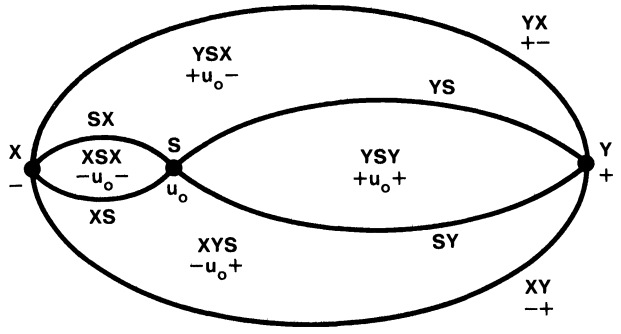

FIG. 8

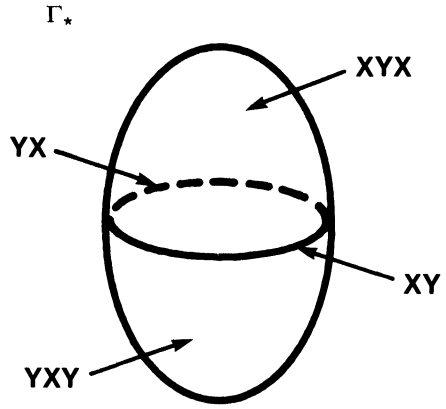

FIG. 9

with $\partial\Gamma_* = \Gamma_{-+-} \cup \Gamma_{+-+}$, i.e., all bang-bang trajectories with at most two switchings (see Fig. 9).

The surface Γ^* consists of bang-bang trajectories analogous to the bang-bang singular case in dimension four. Now

$$\tilde{\Gamma}^- = \{0 \exp(s_1 X) \exp(s_2 Y) \exp(s_3 X) \exp(s_4 Y) \colon s_i \geqq 0,\ s_1 + s_2 + s_3 + s_4 = 1,\ s_1 \leqq s_3,\ s_4 \leqq s_2\text{-conjugate point relations}\},$$

$$\tilde{\Gamma}^+ = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X) \colon t_i \geqq 0,\ t_1 + t_2 + t_3 + t_4 = 1,\ t_1 \leqq t_3,\ t_4 \leqq t_2\text{-conjugate point relations}\}.$$

$\tilde{\Gamma}^-$ and $\tilde{\Gamma}^+$ intersect in a two-dimensional surface $\hat{\Gamma}$, which consists of those trajectories for which

$$(s_1 + s_3)^2 - (s_1 + s_3) + 2 s_2 s_3 = 0,$$

respectively,

$$(t_1 + t_3)^2 - (t_1 + t_3) + 2 t_2 t_3 = 0.$$

The intersection is transversal except at those points that lie on the relative boundary of $\tilde{\Gamma}^-$ or $\tilde{\Gamma}^+$. These points are again characterized by the conjugate point relation

$$\hat{\Gamma} \cap \Gamma_{-+-} = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X) \colon t_1 = 0,\ t_4 = t_2\},$$

$$\hat{\Gamma} \cap \Gamma_{+-+} = \{0 \exp(t_1 Y) \exp(t_2 X) \exp(t_3 Y) \exp(t_4 X) \colon t_1 = t_3,\ t_4 = 0\}.$$

We define Γ^- (respectively, Γ^+) as the component of $\tilde{\Gamma}^-$ ($\tilde{\Gamma}^+$) containing the YX-curve $= \{0 \exp(s_2 Y) \exp(s_3 X) \colon s_i \geqq 0,\ s_2 + s_3 = 1\}$ (respectively, the XY-curve) in its boundary. Then $\Gamma^* := \Gamma^- \cup \Gamma^+$ consists precisely of those points that maximize x_4 on the reachable set. Note that topologically Γ^* also is a stratified sphere with $\partial\Gamma^* = \Gamma_{-+-} \cup \Gamma_{+-+}$, the set of all bang-bang trajectories with at most two switchings (see Fig. 10).

The key fact here is that it is still obvious that $\partial\Gamma^*$ and $\partial\Gamma_*$ match up. They are identical. It is therefore clear that Reach $(0, 1)$ is the set of all points that lie between Γ^* and Γ_*.

It is precisely this simple reasoning that breaks down in the general case. The cause for this lies in the structure of the singular controls. The analysis of the bang-bang trajectories carries over to the general case with only one minor change in the structure. Whereas in the free nilpotent system the two curves $\hat{\Gamma} \cap \Gamma_{+-+}$ and $\hat{\Gamma} \cap \Gamma_{-+-}$ both have points corresponding to the X- and Y-trajectories as endpoints, this need no longer be true: $\hat{\Gamma} \cap \Gamma_{+-+}$ is a curve starting at $0 \exp(1 \cdot Y)$ but which in general no longer ends in $0 \exp(1 \cdot X)$ but rather on a point in the XY-curve (respectively, YX-curve). This distortion is due to the presence of fourth-order brackets. One possible case is depicted in Fig. 11.

Still the relative boundary of Γ^* consists of all bang-bang trajectories with at most two switchings. The structure breaks down in the analysis of the singular surface Γ_* for u near ± 1. The reason is that in the presence of fourth-order brackets the singular controls are no longer constant, and thus the analogue of Γ_{*,u_0} for $u_0 = -1$ does not reduce to Γ_{+-+}, i.e., to bang-bang trajectories with two switchings. For instance, it may not be at all possible to start a singular control with $u_0 = -1$. This is the case if $\dot{u} < 0$ at $u_0 = -1$, which happens under generic assumptions on fourth-order brackets.

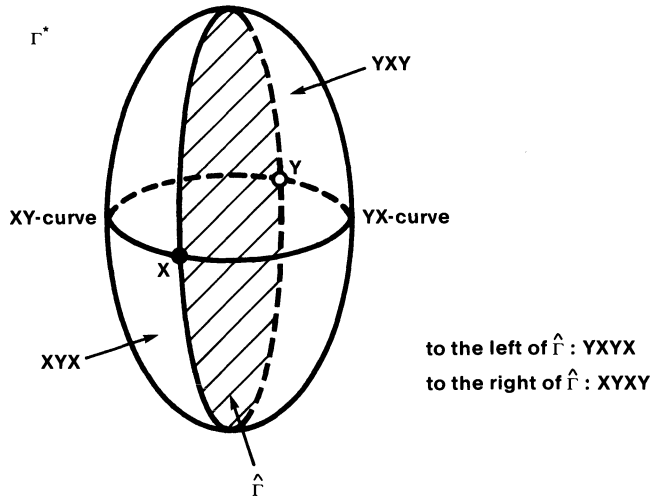

FIG. 10

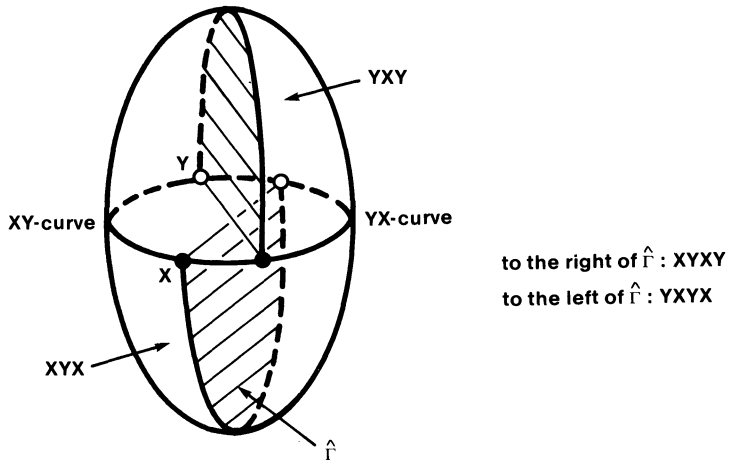

FIG. 11

For the same reason, singular controls with u_0 close to ± 1 may have to be terminated when they become one in absolute value. If the singular control becomes saturated (i.e., hits the constraint and cannot be continued) then this determines the subsequent structure of the trajectory and it is easy to see that concatenations such as *BSBB* or *BBSB*, which are not present in the free nilpotent system, come into play. Therefore Γ_* has trajectories in its relative boundary that contain singular arcs. The main challenge in applying our technique to higher dimensions seems to be finding a way to decide whether structurally different trajectories, such as a bang-bang trajectory, and a concatenation that contains a singular arc steer a system to the same point. Once $\partial\Gamma^*$ and $\partial\Gamma_*$ can be identified, it is clear that the set they enclose is the small-time reachable set.

Note, however, that this structural instability only happens near $\Gamma_{*,-1}$ and $\Gamma_{*,+1}$. The structure of most of the trajectories in the boundary is still the same as in the free nilpotent systems. And it is intuitively clear that the structure of the exceptional trajectories will come up in a higher-dimensional nilpotent system. Therefore, in our

view, the study of the structure of the reachable sets for nilpotent systems will be the key to the general problem.

6. Summary. We have described an approach to determining the qualitative structure of the small-time reachable set in a nondegenerate situation. It is a nontrivial extension of a construction done by Lobry in dimension three. In dimension four we succeed completely in determining the small-time reachable set. For higher dimensions obstacles still have to be overcome. However, they do not lie in the general structure of our approach, but in the fact that too little is known about the structure of extremal trajectories in higher dimensions. For instance, in the five-dimensional case, what is the precise structure of extremal trajectories that contain a saturated singular arc? For dimensions six and beyond, the crucial new ingredient appears to be the incorporation of chattering arcs, another structure of extremal trajectories about which little is still known.

REFERENCES

[1] V. G. BOLTYANSKY, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control, 4 (1966), pp. 326–361.

[2] A. BRESSAN, *Directional convexity and finite optimality conditions*, J. Math. Anal. Appl., 125 (1987), pp. 234–246.

[3] ———, *Local asymptotic approximation of nonlinear control systems*, Tech. Report, University of Wisconsin, Madison, WI, 1984.

[4] ———, *The generic local time-optimal stabilizing controls in dimension* 3, SIAM J. Control Optim., 24 (1986), pp. 177–190.

[5] P. BRUNOVSKY, *Existence of regular synthesis for general problems*, J. Differential Equations, 38 (1980), pp. 317–343.

[6] C. BYRNES AND P. CROUCH, *Local accessibility, local reachability, and representations of compact groups*, Math. Systems Theory, 19 (1986), pp. 43–65.

[7] P. E. CROUCH AND P. C. COLLINGWOOD, *The observation space and realizations of finite Volterra series*, SIAM J. Control Optim., 25 (1987), pp. 316–333.

[8] P. E. CROUCH, *Solvable approximations to control systems*, SIAM J. Control Optim., 22 (1984), pp. 40–45.

[9] W. H. FLEMING AND R. W. RISHEL, *Deterministic and stochastic optimal control*, Applications of Mathematics, Vol. 1, Springer-Verlag, New York, 1975.

[10] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, SIAM J. Control Optim., 16 (1978), pp. 715–727.

[11] N. JACOBSON, *Lie Algebras*, Dover, New York, 1979.

[12] A. KRENER, *Local approximation of control systems*, J. Differential Equations, 19 (1975), pp. 125–133.

[13] ———, *The higher order maximum principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256–293.

[14] C. LOBRY, *Contrôlabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573–605.

[15] L. P. ROTHSCHILD AND E. M. STEIN, *Hypoelliptic differential operators and nilpotent groups*, Acta Math., 137 (1977), pp. 247–320.

[16] H. SCHÄTTLER, *On the local structure of time-optimal bang-bang trajectories in* $\mathbb{R}^3$, SIAM J. Control Optim., 26 (1988), pp. 186–204.

[17] ———, *The local structure of time-optimal trajectories in dimension three under generic conditions*, SIAM J. Control Optim., 26 (1988), pp. 899–918.

[18] H. SUSSMANN, *Analytic stratifications and control theory*, in Proc. International Congress of Mathematics, Helsinki, 1978, pp. 865–871.

[19] ———, *A bang-bang theorem with bounds on the number of switchings*, SIAM J. Control Optim., 17 (1979), pp. 629–651.

[20] ———, *Lie-Volterra expansion for nonlinear systems*, in Mathematical Theory of Networks and Systems, P. Fuhrman, ed., Lecture Notes in Control and Information Science, 58, Springer-Verlag, Berlin, 1984, pp. 822–828.

[21] ———, *Lie brackets and real analyticity in control theory*, in Mathematical Control Theory, Banach Center Publications, Vol. 14, Warsaw, 1984, pp. 515–542.

[22] H. SUSSMANN, *Envelopes, conjugate points and optimal bang-bang extremals*, in Proc. 1985 Paris Conference on Nonlinear Systems, M. Fliess, M. Harewinkel, eds., D. Reidel, Dordrecht, the Netherlands, 1986.

[23] ———, *The structure of time-optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 856–905.

[24] ———, *Regular synthesis for time-optimal control of single-input analytic systems in the plane*, SIAM J. Control Optim., 25 (1987), pp. 1145–1162.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 148-169, January 1989

009

SPECTRAL ASSIGNABILITY FOR DISTRIBUTED PARAMETER SYSTEMS WITH UNBOUNDED SCALAR CONTROL*

RICHARD REBARBER†

Abstract. This article studies a class of control systems in a Hilbert space H given by $\dot{x}(t) = Ax(t) + bu(t)$, where A generates a holomorphic semigroup on H, $u(t)$ is a scalar control, and the control input b is possibly unbounded. Many systems with boundary or point control can be represented in this form. The author considers the question of what eigenvalues $\{\alpha_k\}_{k\in I}$ the closed-loop system can have when $u(t)$ is a feedback control. Shun-Hua Sun's condition on $\{\alpha_k\}_{k\in I}$ [*SIAM J. Control Optim.*, 19 (1981), pp. 730-743] is generalized to the case where b is unbounded but satisfies an admissibility criterion; this condition is generalized further when unbounded feedback elements are allowed. These results are applied to a structurally damped elastic beam with a single point actuator. Similar techniques also prove a spectral assignability result for a damped elastic beam with a moment control force at one end, even though the associated input element is not admissible in the appropriate sense.

Key words. distributed parameter systems, spectral determination, feedback control, holomorphic semigroups, elastic beam

AMS(MOS) subject classifications. 93C20, 93B55, 93B60

1. Introduction.

In this paper we consider the eigenvalue specification problem for a class of distributed parameter systems with scalar control and unbounded input element. We consider the question of what closed-loop eigenvalues can be realized by a feedback control. Sun answers the question completely in [11] in the case where the input element and feedback element are bounded. We will generalize Sun's sufficient condition to the case where the input element is "admissible" in the sense given in [4], but is not necessarily bounded. We also consider a class of systems where the input element is not even admissible. We consider feedback elements that are not necessarily bounded, allowing us to further relax the conditions on the closed-loop eigenvalues. Furthermore, we give a formula for the closed-loop eigenvectors in terms of the open-loop eigenvectors, and a formula for the feedback element in terms of the dual basis to the open-loop eigenvectors. We will apply these results to two examples involving a structurally damped elastic beam.

The systems we consider are of the form

$$\dot{x}(t) = Ax(t) + bu(t), \qquad x(0) = x_0, \tag{1.1}$$

where $x(t) \in H$, which is a Hilbert space with norm $\|\cdot\|$, and $u(t)$ is a scalar control. We assume that A is a closed operator on H which has eigenvectors $\{\varphi_k\}_{k\in I}$ with associated eigenvalues $\{\lambda_k\}_{k\in I}$, and that $\{\varphi_k\}_{k\in I}$ is a Riesz basis for H, i.e., every $x \in H$ can be written as $\sum_{k\in I} x_k\varphi_k$, and there exists m, $M > 0$ such that

$$m\|x\|^2 \leqq \sum_{k\in I} |x_k|^2 \leqq M\|x\|^2.$$

Let the dual of H be represented by a Hilbert space H' (which can be chosen to be H, although this might not be convenient), and let $\langle\cdot,\cdot\rangle$ be the duality relation on $H' \otimes H$. Then the norm in H' is given by $\|y\|_* = \sup\{\langle y, x\rangle / \|x\|\}_{x\in H}$. For $h \in H'$ we

* Received by the editors February 17, 1987; accepted for publication (in revised form) November 25, 1987. This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR-86-0079.

† University of Nebraska, Lincoln, Nebraska 68588.

denote the bounded linear functional $x \to \langle h, x\rangle$ by h^*, and any bounded linear functional on H has such a representation. Let $\{\psi_k\}_{k\in I}$ be the Riesz basis of H' which is biorthogonal to $\{\varphi_k\}_{k\in I}$; i.e., $\langle \psi_k, \varphi_j\rangle = \delta_{j,k}$.

The input element b is a one-dimensional "admissible input element" which can be represented by

$$b = \sum_{k\in I} b_k \varphi_k. \tag{1.2}$$

When b is in H, b_k is given by $\langle \psi_k, b\rangle$, and this will be generalized later. We will use a modification of the definition of admissible input given by Ho and Russell in [5], which we state in the next section. Many systems with boundary or point control can be put into the form (1.1) with an input element that is admissible but unbounded [4], [7], [10].

In this paper we are interested in the case where the eigenvalues lie between two curves of the form

$$\Gamma_a := \{a \pm ix\, e^{\pm i\sigma} | x \in (0, \infty)\}, \tag{1.3}$$

where $\sigma \in [0, \pi/2)$, and σ is fixed for the remainder of this paper. If $\sigma > 0$ A generates a holomorphic semigroup of operators $S(t)$ for t in

$$\tilde{\Omega} := \{z \in \mathbb{C} | |\arg(z)| < \sigma\}, \tag{1.4}$$

and $S(t)$ is strongly continuous on $\bar{\tilde{\Omega}}$. If $\sigma = 0$, A generates a group of operators for $t \in (-\infty, \infty)$. The semigroup (or group) generated by A is given by

$$S(t)\left(\sum_{k\in I} x_k \varphi_k\right) = \sum_{k\in I} x_k\, e^{\lambda_k t} \varphi_k. \tag{1.5}$$

In this paper we consider the following problem for the system (1.1).

Eigenvalue Specification Problem. Given a set of complex numbers $\{\alpha_k\}_{k\in I}$, can a linear functional h^*: $\mathcal{D}(h^*) \subset H \to \mathbb{R}$ be found so that the feedback control $u(t) = h^* x(t)$ leads to a closed-loop system which has eigenvalues at $\{\alpha_k\}_{k\in I}$? If possible, find h^* and the closed-loop eigenvectors $\{\chi_k\}_{k\in I}$.

In this paper we will consider a class of linear functionals h^* which are "admissible," as described in § 2.

For the problem with bounded input (i.e., $b \in H$) and bounded feedback control (i.e., $h \in H'$) the existence question is answered completely for some of the systems described above by Sun [11]. He proves the following necessary and sufficient condition that there exists a bounded functional h^*: $H \to \mathbb{R}$ such that $A + bh^*$ has eigenvalues at $\{\alpha\}_{k\in I}$:

$$\sum_{k\in I} |(\alpha_k - \lambda_k)/b_k|^2 < \infty, \tag{1.6}$$

when $b_j \neq 0$ for all $j \in I$. The restriction that b must be a bounded input rules out many interesting cases, including boundary control. Furthermore, this result is restricted to bounded feedback elements h^*, which is why we cannot move the closed-loop eigenvalues uniformly away from the original eigenvalues. In [11], no formulas are given for the eigenvectors of the closed-loop operator, although they can be found by solving an infinite-dimensional linear system of equations.

Sun's result, like most spectral determination results [2], [9], [12], and this paper, requires a spacing condition on $\{\lambda_k\}_{k\in I}$, so that the eigenvalues are not too close to each other asymptotically (see condition (6) in Definition 3 in the Appendix). In [4], Ho also considers systems with an admissible input element and describes a different

kind of generalization of Sun's condition, where the spacing conditions are relaxed and (1.6) is modified accordingly.

In [2], [7], [9], and [12], eigenvalue specification is studied using canonical forms for (1.1). In [7], [9], and [12], $\{\lambda_k\}_{k\in I}$ is the zero set of a *cardinal function*, which is an infinite-dimensional generalization of a characteristic polynomial of a finite-dimensional system. In [9] and [12] the cardinal functions are entire functions. In this paper we will be using the definition of a cardinal function given in [7], where the cardinal function is a meromorphic function, which is applicable to more cases than entire cardinal functions. This definition is given in the Appendix to this paper. We need to consider meromorphic cardinal functions for some systems because there is no entire function with zeros at $\{\lambda_k\}_{k\in I}$ which have appropriate growth conditions in the right half of the complex plane. Classes of cardinal functions are constructed in [8]. For the remainder of the paper, we will insist on the following assumption.

Assumption A. $\{\lambda_k\}_{k\in I}$ is the zero set of some cardinal function p.

In [7] we used canonical forms to consider the cases where $\{\alpha_k\}_{k\in I}$ is the zero set of a cardinal function with the same poles as p and either

(I) b is a bounded input, $b_k \neq 0$ for all $k \in I$, and (1.6) holds, or

(II) There exists $m, M > 0$ such that

$$m < b_k < M. \tag{1.7}$$

(Note that condition (II) includes no other restrictions on $\{\alpha_k\}_{k\in I}$.)

In both cases (I) and (II) we find a formula for an h^* such that $A+bh^*$ has eigenvalues at $\{\alpha_k\}_{k\in I}$, and give a formula for the closed-loop eigenvectors. In case (I), h^* is a bounded feedback element, and in case (II), h^* is an admissible feedback element, as defined in § 2. We also use the canonical form in [7] to prove a control spillover results for a finite-dimensional approximation to the infinite-dimensional feedback control in cases (I) and (II). In this paper we will generalize those formulas and study their applicability in detail. Those generalizations follow: let

$$J := \{j \in I \,|\, b_j \neq 0\}, \qquad K := \{k \in I \,|\, b_k = 0\}, \tag{1.8}$$

and assume that

$$\alpha_k = \lambda_k \quad \text{for } k \in K. \tag{1.9}$$

The closed-loop eigenvectors $\{\chi_k\}_{k\in I}$ associated with $\{\alpha_k\}_{k\in I}$ are

$$\begin{aligned} \chi_k &= (1/p'(\lambda_k)b_k) \sum_{j\in J} [p(\alpha_k)b_j/(\alpha_k - \lambda_j)]\varphi_j \quad \text{for } k \in J, \\ \chi_k &= \varphi_k \quad \text{for } k \in K. \end{aligned} \tag{1.10}$$

Under certain conditions, we will show that these form a Riesz basis for H. The basis for H' which is biorthogonal to $\{\chi_k\}_{k\in I}$ is defined as follows:

$$\begin{aligned} h_j &= b_j p'(\lambda_j) \sum_{k\in J} \{q(\lambda_k)/[q'(\alpha_j)(\lambda_k - \alpha_j)b_k p'(\lambda_k)]\}\psi_k \quad \text{for } j \in J, \\ h_j &= \psi_j \quad \text{for } j \in K, \end{aligned} \tag{1.11}$$

where q is a cardinal function with zeros at $\{\alpha_k\}_{k\in I}$ and the same poles as p. The feedback element will be given by

$$h^* = \sum_{j\in J} p(\alpha_j)h_j/b_j p'(\lambda_j). \tag{1.12}$$

In this paper we will work with these equations directly, without reference to canonical forms. We will show that these formulas solve the Eigenvalue Specification

Problem under conditions much less restrictive than conditions (I) and (II). In the next section we give the definitions of "admissible input element" and "admissible feedback element" which we will be using. In § 3 we will state and prove control results related to formulas (1.10)–(1.12) for systems with a general admissible input, and for systems with a class of input elements which are not even admissible. In § 4 we will describe in detail an example of a damped elastic beam with two kinds of controls. We apply our results to an example where the control is a single point actuator, and we study how the placement of the actuator affects the kind of eigenvalue specification results we obtain. In the other example, the control is a moment force at one end of the beam. In this case the input element is not even admissible, but we show that we can still use the above formulas to solve the Eigenvalue Specification Problem.

2. Definitions.

Admissible input element. We start by identifying the set of admissible controls $u(t)$: Since $S(t)$ generates a holomorphic semigroup, we will consider the system (1.1) along any ray

$$l_\theta = \{x e^{i\theta} \mid x \in [0, \infty)\} \tag{2.1}$$

for $\theta \in [-\sigma, \sigma]$. The set of admissible controls will be

$$\mathcal{U} := \bigcup_{-\sigma \leqq \theta \leqq \sigma} L^2_{\text{loc}}[l_\theta].$$

We will be using a modification of the definition of admissible input element given in Ho and Russell [5], which is the same modification used by Ho in [4]. To motivate this, we must first decide which space b should belong to.

When $z_0 \in \mathcal{D}(A)$, $S(t)z_0$ is an element of H for all $t \in \tilde{\tilde{\Omega}}$ and is the solution in H of

$$\dot{z}(t) = Az(t), \qquad z(0) = z_0. \tag{2.2}$$

If $z_0 \in H$, $z(t) = S(t)z_0$ is an element of H, and is the solution of (2.2) in a generalized sense, but (2.2) is no longer an equation in H. A can be extended to an operator on $\mathcal{H} := \mathcal{D}(A^*)'$ as follows:

$$\hat{A}\colon H \to \mathcal{H}\colon \quad \langle \hat{A}z, \eta \rangle := \langle z, A^*\eta \rangle \quad \forall \eta \in \mathcal{D}(A^*), \tag{2.3}$$

where A^* is the adjoint of A when the inner product $\langle \cdot, \cdot \rangle$ is used. In this context we think of $\mathcal{D}(A^*)$ as a Hilbert space with the graph norm, and $\mathcal{D}(A^*) \subset H'$. We think of $\mathcal{H}$ as a Hilbert space with the norm induced by the duality pairing $\langle \cdot, \cdot \rangle$. Therefore, (2.2) is an equation in $\mathcal{H}$. If we want (1.1) to be an equation in $\mathcal{H}$, we must at least require that b is an element of $\mathcal{H}$, i.e., b is a bounded linear functional on the Hilbert space $\mathcal{D}(A^*)$.

Not all $b \in \mathcal{H}$ will lead to a dynamical system in H. If $b \in H$, $t \in l_\theta$, and $u(t) \in L^2_{\text{loc}}[l_\theta]$, the solution of (1.1) with $x_0 = 0$ is

$$B(t)u = \int_{[0,t]} (S(t-s)b)u(s)\, ds,$$

where $[0, t]$ is the straight line segment from the origin to t. A generalization of this to the case where $b \in \mathcal{H}$ is

$$\langle B(t)u, y \rangle = \int_{[0,t]} \langle b, S(t-s)^* y \rangle u(s)\, ds \quad \text{for every } y \in \mathcal{D}(A^*). \tag{2.4}$$

DEFINITION 1. $b \in \mathcal{H}$ is an admissible input element on l_θ if $B(t)$ given by (2.4) is a strongly continuous family of bounded operators from $L^2_{\text{loc}}[l_\theta]$ into H. b is admissible on the wedge $\tilde{\tilde{\Omega}}$ (cf. (1.3)) if it is admissible on l_θ for all $\theta \in [-\sigma, \sigma]$.

If b is admissible on l_θ, $x(t)=S(t)x_0+B(t)u$ is a generalized solution of (1.1) in the sense that $x(t)\in H$, and $x(0)=x_0$, $\dot{x}(t)\in\mathscr{H}$, and $\dot{x}(t)=\hat{A}x(t)+bu(t)$ in $\mathscr{H}$ for all $t\in l_\theta$.

If $x_0=\sum_{k\in I}x_{0,k}\varphi_k$ and b is given by (1.2), the generalized solution of (1.1) is

$$x(t)=\sum_{k\in I}\left(x_{0,k}\,e^{\lambda_k t}+b_k\int_{[0,t]}e^{\lambda_k(t-s)}u(s)\,ds\right)\varphi_k. \tag{2.5}$$

In this case, b is admissible on l_θ if and only if $x(t)$ given by (2.5) is an element of H whenever $u\in L^2_{\mathrm{loc}}[l_\theta]$.

An easily checked sufficiency condition for admissibility is given in [5]. It involves checking whether a certain measure is a Carleson measure. For instance, using this method it is easy to see that, for the systems under consideration in this paper, an input element b with $\{b_k\}_{k\in I}\in l_\infty$ is admissible.

Admissible feedback elements. Let $h^*\colon H\to\mathbb{R}$ be a linear functional with domain $\mathscr{D}(h^*)$. For any $x\in\mathscr{D}(h^*)$, we can define $\hat{A}x+bh^*x$ (cf. (2.3)) as an element of $\mathscr{H}$. We will define the domain of $A+bh^*$ as

$$\mathscr{D}(A+bh^*):=\{x\in\mathscr{D}(h^*)\,|\,\hat{A}x+bh^*x\in H\}. \tag{2.6}$$

For $x\in\mathscr{D}(A+bh^*)$, we define

$$(A+bh^*)x=\hat{A}x+bh^*x. \tag{2.7}$$

DEFINITION 2. h^* is an admissible feedback element on l_θ if

(1) $\mathscr{D}(A+bh^*)$ is dense in H.

(2) There exists A_h, an extension of $A+bh^*$, such that A_h is the infinitesimal generator of a strongly continuous semigroup $S_h(t)$ for $t\in l_\theta$.

(3) For all $x\in H$ and $T\in l_\theta$,

$$\|h^*S_h(t)x\|_{L^2[0,T]}<\infty.$$

h^* is an admissible feedback element on $\tilde{\Omega}$ if h^* is admissible on l_θ for all $\theta\in[-\sigma,\sigma]$.

The third condition in this definition guarantees that the feedback control $u(t)=h^*x(t)$ is admissible. The other two conditions seem to be minimum requirements for the closed-loop operator $A+bh^*$ to be useful. Unbounded feedback elements are considered in [3], where conditions are given under which an unbounded feedback leads to a closed-loop operator with the properties given in Definition 2. In general it will be hard to verify the conditions in Definition 2. In this paper we will verify that h^* is an admissible feedback element by determining the basis properties of the closed-loop eigenvectors χ_k given in (1.12). For instance, it is easy to prove the following proposition, whose proof we omit.

PROPOSITION 1. *Suppose* $\{\chi_k\}_{k\in I}$ *forms a Riesz basis for* H:

$$(A+bh^*)\chi_k=\alpha_k\chi_k, \tag{2.8}$$

and

$$\left\{\sum_{k\in I}x_k\chi_k\,|\,\sum_{k\in I}|x_k\alpha_k|^2<\infty\right\}\subset\mathscr{D}(h^*). \tag{2.9}$$

Then

$$\mathscr{D}(A+bh^*)=\left\{\sum_{k\in I}x_k\chi_k\,|\,\sum_{k\in I}|x_k\alpha_k|^2<\infty\right\}, \tag{2.10}$$

and $A+bh^*$ is the infinitesimal generator of the semigroup

$$S(t)\left(\sum_{k\in I} x_k\chi_k\right) = \sum_{k\in I} x_k e^{\alpha_k t}\chi_k, \tag{2.11}$$

which is holomorphic on $\tilde{\Omega}$ and strongly continuous on $\bar{\tilde{\Omega}}$.

3. Control results. We will begin this section by stating the main results of this paper. We first need some notation. Suppose $\{\alpha_k\}_{k\in I}$ is the set we would like to realize as the closed-loop eigenvalues and $\{\alpha_k\}$ is the zero set of a cardinal function. Let m be such that

$$|\alpha_j-\alpha_k|\geqq m|j-k| \quad \text{and} \quad |\lambda_j-\lambda_k|\geqq m|j-k|, \tag{3.1}$$

where the existence of such an m is guaranteed by condition (6), the definition of a cardinal function, in Definition 3 in the Appendix. Fix m for the remainder of this section. Let

$$\begin{aligned} &j_k\in J_k := \{j\in I \,|\, |\alpha_j-\lambda_k| \text{ is minimized}\},\\ &\mathcal{J} := \{k\in I \,|\, |\alpha_{j_k}-\lambda_k|<m/2\}, \end{aligned} \tag{3.2}$$

$$\begin{aligned} &I_k = I \quad \text{for } k\notin\mathcal{J},\\ &I_k = I\backslash\{j_k\} \quad \text{for } k\in\mathcal{J}. \end{aligned} \tag{3.3}$$

THEOREM 2. *Suppose $\{\alpha_k\}_{k\in I}$ is the zero set of a cardinal function q which has the same poles as p, b is admissible on $\bar{\tilde{\Omega}}$, and*

$$\{(\alpha_k-\lambda_k)/b_k\}_{k\in I}\in l_\infty, \tag{3.4}$$

$$\left\{(\alpha_k-\lambda_k)\sum_{j\in\mathcal{J}}|1/(\alpha_k-\lambda_j)|^2\right\}_{k\in I}\in l_\infty. \tag{3.5}$$

Then h^* given by (1.12) *is an admissible feedback element on $\bar{\tilde{\Omega}}$ with domain*

$$\mathcal{D}(h^*) = \left\{\sum_{k\in I} x_k\chi_k \,\Big|\, \sum_{k\in I}|x_k|<\infty\right\}, \tag{3.6}$$

and $A+bh^*$ has eigenvalues at $\{\alpha_k\}_{k\in I}$ and eigenvectors $\{\chi_k\}_{k\in I}$ given by (1.10) *which form a Riesz basis for H.*

In some cases (3.5) is easy to verify—for instance, we will show that (3.5) is true if $\{\alpha_k-\lambda_k\}_{k\in I}\in l_\infty$.

The next theorem is a more direct generalization of Sun's result in [11].

THEOREM 3. *Suppose $\{\alpha_k\}_{k\in I}$ is the zero set of a cardinal function q which has the same poles as p, b is admissible on $\bar{\tilde{\Omega}}$, and*

$$\{(\alpha_j-\lambda_k)/b_k\}_{k\in I}\in l_2. \tag{3.7}$$

Then h^* is a bounded feedback element, and $A+bh^*$ has eigenvalues at $\{\alpha_k\}_{k\in I}$ and associated eigenvectors $\{\chi_k\}_{k\in I}$.

We can use similar methods to prove eigenvalue specification results for some systems which do not satisfy the hypotheses of Theorem 2 or Theorem 3.

THEOREM 4. *Let $b_k=\beta_k k$ and $\lambda_k=\eta_k k^2$, where*

$$M_1\geqq|\beta_k|\geqq m_1>0 \quad \text{and} \quad M_2\geqq|\eta_k|\geqq m_2>0 \tag{3.8}$$

for some m_1, M_1, m_2, and M_2. (In [8] *it is shown that there is a cardinal function p with zeros at $\{\lambda_k\}_{k\in I}$.) Suppose that $\{(\alpha_k-\lambda_k)/b_k\}_{k\in I}\in l_\infty$ and there exists a cardinal function q with the same poles as p. Then $\{\chi_k\}_{k\in I}$ given by* (1.10) *is a Riesz basis for H, h^* given by* (1.12) *with domain* (3.6) *is an admissible feedback element, and $A+bh^*$ has eigenvectors $\{\chi_k\}_{k\in I}$ and eigenvalues $\{\alpha_k\}_{k\in I}$.*

We will show that under the hypotheses of Theorem 4, the input element b is not admissible on $\bar{\Omega}$. It is usually difficult to analyze systems like this because the solution is not guaranteed to be in H when u is an admissible control.

The following results from [7] will be needed in this section, so they are stated as a lemma.

LEMMA 5. *Suppose p and q are both cardinal functions with the same poles, and let p have zeros at $\{\lambda_k\}_{k\in I}$ and q have zeros at $\{\alpha_k\}_{k\in I}$. Then*

$$\langle h_j, \chi_k\rangle_* = \delta_{j,k} \quad \text{for all } j, k \in J \quad (\text{cf. } (1.8)), \tag{3.9}$$

$$\varphi_k = (1/p'(\lambda_k)b_k)\sum_{j\in J}[q(\lambda_k)p'(\lambda_j)b_j/(\lambda_k-\alpha_j)q'(\alpha_j)]\chi_j, \tag{3.10}$$

$$\psi_k = \sum_{j\in J}[b_k p(\alpha_j)/(\alpha_j-\lambda_k)b_j]h_j, \tag{3.11}$$

$$\begin{aligned}&\{|p(\alpha_j)/(\lambda_k-\alpha_j)|\}_{j,k\in I}\in l_\infty,\\ &\{|q(\lambda_j)/(\alpha_k-\lambda_j)|\}_{j,k\in I}\in l_\infty,\end{aligned} \tag{3.12}$$

$$m \leqq p'(\lambda_k) \leqq M \quad \text{for some } m, M>0. \tag{3.13}$$

Formula (3.12) *and the second inequality in* (3.13) *are consequences of the definition of a cardinal function, and their proofs are given in the Appendix. The other results are proved in* [7].

Note that the index set in (3.10) and (3.11) is J, given by (1.8), and not I. Our first task is to show that it suffices to assume that $b_k \neq 0$ for all $k \in I$, i.e., that $J = I$.

LEMMA 6. *Suppose $\{\chi_k\}_{k\in J}$ given by* (1.10) *is a Riesz basis for*

$$H_J := \left\{\sum_{k\in J} x_k\varphi_k \,|\, \{x_k\}\in l_2\right\}.$$

Then $\{\chi_k\}_{k\in I}$ is a Riesz basis for H, and $\{\chi_k\}_{k\in I}$ is biorthogonal to $\{h_j\}_{j\in I}$ given by (1.11), *which is a Riesz basis for H'.*

Proof. Using (3.11), we can see that $\{h_j\}_{j\in J}$ is a basis for

$$H'_J := \left\{\sum_{k\in J} x_k\psi_k \,|\, \{x_k\}\in l_2\right\}.$$

It follows from (3.9) and the hypotheses that $\{h_j\}_{j\in J}$ is a Riesz basis for H'_J. $\{\chi_k\}_{k\in K}$ (cf. (1.8), (1.10)) is obviously a Riesz basis for

$$H_K := \left\{\sum_{k\in K} x_k\varphi_k \,|\, \{x_k\}\in l_2\right\}$$

and $\{h_j\}_{j\in K}$ (cf. (1.11)) is obviously a Riesz basis for

$$H'_K := \left\{\sum_{k\in K} x_k\varphi_k \,|\, \{x\}\in l_2\right\}.$$

If $x = \sum_{k\in I} x_k\varphi_k \in H$, x can be written as $x_1 + x_2$, where $x_1 \in H_J$ and $x_2 \in H_K$. The above statements then imply that

$$\|x\|^2 \leqq M\sum_{k\in I}|x_k|^2 \quad \text{for some } M>0. \tag{3.14}$$

Similarly, we can show that

$$\left\|\sum_{k\in I} y_k h_k\right\|^2 \leqq M\sum_{k\in I}|y_k|^2 \quad \text{for some } M>0. \tag{3.15}$$

We now prove that $\{\chi_k\}_{k\in I}$ is biorthogonal to $\{h_j\}_{j\in I}$: for j and k both in J, or j and k both in K, it is clear that $\langle h_k, \chi_j\rangle = \delta_{j,k}$. Suppose $j\in J$ and $k\in K$. Then $b_k = 0$, and $h_k = \psi_k$; therefore,

$$\langle h_k, \chi_j\rangle = (1/p'(\lambda_j)b_j) \sum_{j\in J} [p(\alpha_j)b_j/(\alpha_j - \lambda_i)]\langle \psi_k, \varphi_i\rangle = 0,$$

since k does not belong to J. Similarly, $\langle h_k, \chi_j\rangle = 0$ when $k\in J$ and $j\in K$; thus we have proved the biorthogonality.

Now note that

$$\sum_{k\in J} |x_k|^2 = \left\langle \sum_{k\in J} x_k h_k, \sum_{k\in I} x_k\chi_k \right\rangle \leqq \left\| \sum_{k\in J} x_k h_k \right\| \left\| \sum_{k\in I} x_k\chi_k \right\| \leqq M \left(\sum_{k\in I} |x_k|^2 \right)^{1/2} \left\| \sum_{k\in I} x_k\chi_k \right\|$$

for some M, using (3.15). Combining this with (3.14), we see that $\{\chi_k\}_{k\in I}$ is a Riesz basis for H, and so $\{h_j\}_{j\in I}$ is a Riesz basis for H'. □

LEMMA 7. *If* $\chi_k \in \mathcal{D}(h^*)$, *then* $(A + bh^*)\chi_k = \alpha_k\chi_k$.

Proof. First note that

$$\hat{A}\chi_k = \sum_{j\in I} \{p(\alpha_k)b_j\lambda_j/[p'(\lambda_k)b_k(\alpha_k - \lambda_j)]\}\varphi_j \quad \text{for } k\in J \tag{3.16}$$

and (since $\chi_k = \varphi_k$ and $\alpha_k = \lambda_k$ for $k\in K$)

$$\hat{A}\chi_k = \alpha_k\chi_k \quad \text{for } k\in K. \tag{3.17}$$

Also,

$$bh^*\chi_k = \sum_{j\in I} [p(\alpha_k)b_j/b_k p'(\lambda_k)]\varphi_j \quad \text{for } k\in J \tag{3.18}$$

and

$$h^*\chi_k = 0 \quad \text{for } k\in K. \tag{3.19}$$

Putting (3.16) and (3.18) together, we get $(A + bh^*)\chi_k = \hat{A}\chi_k + bh^*\chi_k = (1/p'(\lambda_k)b_k)$ $\sum_{j\in J} [\alpha_k p(\alpha_k)b_j/(\alpha_k - \lambda_j)]\varphi_j = \alpha_k\chi_k$, for $k\in J$. Putting (3.17) and (3.19) together, we also get $(A + bh^*)\chi_k = \alpha_k\chi_k$ for $k\in K$. □

The theorems in this section will be proved by showing that $\{\chi_k\}_{k\in I}$ is a Riesz basis for H and applying Proposition 1. Because of Lemma 6, to do this it suffices to assume that the index set K is empty, i.e., $b_k \neq 0$ for all $k\in I$.

We need to verify two inequalities in order to show that $\{\chi_k\}_{k\in I}$ is a Riesz basis. Suppose $\{x_k\}\in l_2$. Then

$$\sum_{k\in I} x_k\chi_k = \sum_{k\in I} x_k(1/p'(\lambda_k)b_k) \sum_{j\in I} [p(\alpha_k)b_j/(\alpha_k - \lambda_j)]\varphi_j = \sum_{j\in I} y_j\varphi_j, \tag{3.20}$$

where

$$y_j = \sum_{k\in I} x_k p(\alpha_k)b_j/[b_k p'(\lambda_k)(\alpha_k - \lambda_j)]. \tag{3.21}$$

Since $\{\varphi_k\}_{k\in I}$ is a Riesz basis of H, to show that

$$\left\| \sum_{k\in I} x_k\chi_k \right\|^2 \leqq M \sum_{k\in I} |x_k|^2 \tag{3.22}$$

is true for some M, we need to show that

$$\sum_{k\in I} |y_k|^2 \leqq M \sum_{k\in I} |x_k|^2 \tag{3.23}$$

is true for some M when y_k is given by (3.21). In this case $\{\chi_k\}_{k\in I}$ is said to be *uniformly* l_2*-convergent.*

If $\{\alpha_k\}_{k\in I}$ is the zero set of a cardinal function with the same poles as p, then (3.10) is true; therefore, we can write

$$\sum_{k\in I} x_k\varphi_k = \sum_{k\in I} x_k(1/p'(\lambda_k)b_k)\sum_{j\in I}[q(\lambda_k)p'(\lambda_j)b_j/(\lambda_k-\alpha_j)q'(\alpha_j)]\chi_j = \sum_{j\in I} y_j\chi_j \tag{3.24}$$

where

$$y_j = \sum_{k\in I} x_k q(\lambda_k)b_j p'(\lambda_j)/[b_k p'(\lambda_k)q'(\alpha_j)(\lambda_k-\alpha_j)]. \tag{3.25}$$

In this case $\{\chi_k\}_{k\in I}$ is a basis for H. Furthermore, the existence of a biorthogonal set (Lemma 5) implies that the χ_k are strongly independent. If $\{\chi_k\}_{k\in I}$ is uniformly l_2-convergent and also *uniformly l_2-independent*, i.e.,

$$m\sum_{k\in I}|x_k|^2 \leqq \left\|\sum_{k\in I} x_k\chi_k\right\|^2 \tag{3.26}$$

for some $m>0$, then $\{\chi_k\}_{k\in I}$ is a Riesz basis. This is true if (3.23) is true for some M when y_j is given by (3.25).

When $\sigma>0$ in (1.3) and A generates a holomorphic semigroup on $\tilde{\Omega}$ that is strongly continuous on $\bar{\tilde{\Omega}}$, the results in this section require that the input element should be admissible on $\bar{\tilde{\Omega}}$, not just on the positive real axis. When A generates a group, we require that the input element is admissible on the positive real axis. The following two lemmas are consequences of admissibility.

LEMMA 8. *Suppose b is admissible on $\bar{\tilde{\Omega}}$, $\{\alpha_k\}_{k\in I}$ satisfies* (3.1), *and we can find α and β such that $\{\alpha_k\}_{k\in I}$ and $\{\lambda_k\}_{k\in I}$ lie between Γ_α and Γ_β (cf.* (1.3)). (*Unlike the rest of the results in this paper, we do not require Assumption* A *for this lemma.*) *Then there exists M such that*

$$\sum_{k\in I}\left|\sum_{j\in I_k} b_k x_j/(\alpha_j-\lambda_k)\right|^2 \leqq M\sum_{j\in I_k}|x_j|^2 \tag{3.27}$$

(*cf.* (3.3)). *If there exists $c>0$ such that*

$$|\alpha_j-\lambda_k|>c \quad \textit{for all } j\in I \textit{ and } k\in I, \tag{3.28}$$

then

$$\sum_{k\in I}\left|\sum_{j\in I} b_k x_j/(\alpha_j-\lambda_k)\right|^2 \leqq M\sum_{j\in I}|x_j|^2. \tag{3.29}$$

Proof. Let m be as in (3.1), so that (cf. (3.2))

$$|\alpha_j-\lambda_k| \geqq |\alpha_{j_k}-\alpha_j|-|\alpha_{j_k}-\lambda_k| \geqq m|j-j_k|-|\alpha_j-\lambda_k|.$$

This implies that

$$|\alpha_j-\lambda_k| \geqq (m/2)|j-j_k|. \tag{3.30}$$

In particular, this implies $|\alpha_j-\lambda_k|\geqq m/2$ for $j\neq j_k$, so that $J_k=\{j_k\}$ if $|\alpha_{j_k}-\lambda_k|<m/2$. Hence (cf. (3.3))

$$|\alpha_j-\lambda_k| \geqq m/2 \quad \text{for } j\in I_k,\ k\in I. \tag{3.31}$$

For future reference, we also note that (3.30) implies that

$$\left\{\sum_{j\in I_k}|1/(\alpha_j-\lambda_k)|^2\right\}_{k\in I} \in l_\infty. \tag{3.32}$$

Let $u(s)=\sum_{j\in I} x_j e^{\alpha_j s}$. We will first assume that the following two conditions hold:

Condition 1. $\{\alpha_k\}_{k\in I}$ is to the left of Γ_0.

Condition 2. There exists $\rho, r\in\mathbb{R}$ such that $\rho<r$, $\{\alpha_k\}_{k\in I}$ is to the left of Γ_ρ (cf. (1.3)) and $\{\lambda_k\}_{k\in I}$ is to the right of Γ_r.

Then

$$\int_{l_\theta} |u(s)|^2\,|ds| \leqq M \sum_{j\in I} |x_j|^2 \tag{3.33}$$

for some M independent of $\theta\in[-\sigma,\sigma]$, by results in [6]. Using (2.5), we see that since b is admissible on $\tilde{\tilde{\Omega}}$,

$$\sum_{k\in I} \left| b_k \int_{[0,t]} e^{\lambda_k(t-s)} u(s)\,ds \right|^2 \leqq M \int_{[0,t]} |u(s)|^2 |ds| \tag{3.34}$$

for some M and all $t\in\tilde{\tilde{\Omega}}$. Let $J_1 := \{k\in I \mid \operatorname{Im}(\lambda_k)>0\}$ and $K_1 := \{k\in I \mid \operatorname{Im}(\lambda_k)\leqq 0\}$. For all $t\in l_{-\sigma}$, the left side of (3.34) is greater than or equal to

$$\sum_{k\in J_1} |e^{\lambda_k t}|^2 \left| b_k \int_{[0,t]} e^{-\lambda_k s} u(s)\,ds \right|^2 \geqq \tilde{m} \sum_{k\in J_1} \left| b_k \int_{[0,t]} e^{-\lambda_k s} u(s)\,ds \right|^2$$

for some $\tilde{m}>0$, since $\operatorname{Re}(\lambda_k t)$ is bounded below for $k\in J_1$ and $t\in l_{-\sigma}$. Let $t\to\infty$ on $l_{-\sigma}$, and the integral on the right side approaches the Laplace transform (as defined in [6]) of u evaluated at λ_k, or $\sum_{j\in I} x_j/(\alpha_j-\lambda_k)$.

Putting this together with (3.33) and (3.34), we see that

$$\sum_{k\in J_1} \left| \sum_{j\in I} b_k x_j/(\alpha_j-\lambda_k) \right|^2 \leqq M \sum_{j\in I} |x_j|^2$$

for some M. Now let $t\in l_\sigma$, and we do the same thing with the index set K_1 replacing J_1, yielding (3.29), as long as Conditions 1 and 2 are satisfied.

Let $a>0$ be large enough so that

$$\{\alpha_k - a\}_{k\in I} \text{ satisfies Conditions 1 and 2.} \tag{3.35}$$

We can verify that

$$\begin{aligned}
\sum_{k\in I} \left| \sum_{j\in I_k} b_k x_j/(\alpha_j-\lambda_k) \right|^2 &= \sum_{k\in I} \Bigg| \sum_{j\in I_k} b_k x_j/(\alpha_j - a - \lambda_k) \\
&\qquad - \sum_{j\in I_k} b_k x_j a/(\alpha_j - a - \lambda_k)(\alpha_j - \lambda_k) \Bigg|^2 \\
&\leqq \sum_{k\in I} \left| \sum_{j\in I_k} b_k x_j/(\alpha_j - a - \lambda_k) \right|^2 \\
&\qquad + \sum_{k\in I} \left| \sum_{j\in I_k} b_k x_j a/(\alpha_j - a - \lambda_k)(\alpha_j - \lambda_k) \right|^2 .
\end{aligned} \tag{3.36}$$

The first sum in the above line is bounded by $M\sum_{k\in I}|x_k|^2$ for some M, because (3.35) implies that (3.29) is true when $\{\alpha_k\}_{k\in I}$ is replaced by $\{\alpha_k - a\}_{k\in I}$. The second sum is less than or equal to

$$\begin{aligned}
&\sum_{k\in I} \sum_{j\in I_k} |b_k x_j a/(\alpha_j - a - \lambda_k)|^2 \left\{ \sum_{j\in I_k} |1/(\alpha_j - \lambda_k)|^2 \right\} \\
&\qquad \leqq \left\{ \sup_{k\in I} \sum_{j\in I_k} |1/(\alpha_j - \lambda_k)|^2 \right\} \sum_{j\in I} |x_j|^2 \left[\sum_{k\in I} |b_k a/(\alpha_j - a - \lambda_k)|^2 \right].
\end{aligned} \tag{3.37}$$

The first term on the right in (3.37) is bounded, from (3.32). Using (3.29) and (3.35), we see that

$$\sum_{k\in I} |b_k/(\alpha_j-\lambda_k-a)|^2 \leqq M \tag{3.38}$$

for some M. Hence (3.37) is bounded by $M\sum_{k\in I}|x_k|^2$ for some M. Combining this with (3.36) yields (3.27).

If (3.28) is true, we can go through the same proof by replacing the summation over $j\in I_k$ in (3.36) and (3.37) with the summation over $j\in I$. This leads to (3.29) and finishes the proof of the lemma. □

We can use Lemma 8 to study the basis properties of $\{\chi_k\}_{k\in I}$.

LEMMA 9. *Suppose b is admissible on $\tilde{\Omega}$, $\{\alpha_k\}_{k\in I}$ satisfies* (3.4), *and we can find α and β such that $\{\alpha_k\}_{k\in I}$ and $\{\lambda_k\}_{k\in I}$ lie between Γ_α and Γ_β. Then* (3.23) *is true with y_j given by* (3.21).

Proof. Using (3.21), we have that

$$\sum_{k\in I}|y_k|^2=\sum_{k\in I}\left|\sum_{j\in I}[x_jp(\alpha_j)b_k]/[b_jp'(\lambda_j)(\alpha_j-\lambda_k)]\right|^2. \tag{3.39}$$

The right side of (3.39) is less than or equal to

$$\sum_{k\notin\mathcal{J}}\left|\sum_{j\in I}[x_jp(\alpha_j)b_k]/[b_jp'(\lambda_j)(\alpha_j-\lambda_k)]\right|^2+\sum_{k\in\mathcal{J}}\left|\sum_{j\in I_k}[x_jp(\alpha_j)b_k]/[b_jp'(\lambda_j)(\alpha_j-\lambda_k)]\right|^2$$

$$+\sum_{k\in\mathcal{J}}|[x_{j_k}p(\alpha_{j_k})b_k]/[b_{j_k}p'(\lambda_{j_k})(\alpha_{j_k}-\lambda_k)]|^2 \tag{3.40}$$

(cf. (3.2), (3.3)). By (3.12), (3.13), and (3.4), $|p(\alpha_j)/b_jp'(\lambda_j)|$ is bounded. The first two sums are bounded by

$$M\sum_{k\notin\mathcal{J}}\left|\sum_{j\in I}x_jb_k/(\alpha_j-\lambda_k)\right|^2+M\sum_{k\in\mathcal{J}}\left|\sum_{j\in I_k}x_jb_k/(\alpha_j-\lambda_k)\right|^2\leqq\tilde{M}\sum_{j\in I}|x_j|^2$$

for some M and $\tilde{M}$, where we use the two results in Lemma 5.

Using (3.4) and (3.12), we see that there exists M such that the third sum in (3.40) is less than or equal to

$$M\sum_{k\in\mathcal{J}}|x_{j_k}b_k/b_{j_k}|^2. \tag{3.41}$$

To get an upper bound on this we need the following lemma.

LEMMA 10. $\{|b_k/b_{j_k}|\}_{k\in\mathcal{J}}$ *is bounded.*

Proof of Lemma 10. If $j_k=k$, then $|b_k/b_{j_k}|=1$. Now assume that $j_k\neq k$ and $k\in\mathcal{J}$. When we let a be as in (3.35), (3.38) is true and implies that

$$|b_k|\leqq M|\alpha_{j_k}-\lambda_k-a|,$$

which is bounded by $M(m/2+a)$ when $k\in\mathcal{J}$ (cf. (3.3)). Using (3.4), we have that

$$|1/b_{j_k}|\leqq\tilde{M}|1/(\alpha_{j_k}-\lambda_{j_k})$$

for some $\tilde{M}$. For $k\in\mathcal{J}$ and $j_k\neq k$, $|\alpha_{j_k}-\lambda_{j_k}|=|\alpha_{j_k}-\lambda_k+\lambda_k-\lambda_{j_k}|\geqq|\lambda_k-\lambda_{j_k}|-|\alpha_{j_k}-\lambda_k|\geqq m|k-j_k|-m/2\geqq m/2$ (cf. (3.1) and (3.31)). Therefore, $|b_k/b_{j_k}|\leqq M\tilde{M}(m/2+a)/(m/2)$ for $k\in\mathcal{J}$ and $j_k\neq k$, so Lemma 10 is true.

This means that (3.40), and hence (3.39), is bounded by $M\sum_{k\in I}|x_k|^2$ for some M, which completes the proof of Lemma 9. □

THEOREM 11. *Suppose* $\{\alpha_k\}_{k\in I}$ *is the zero set of a cardinal function* q *which has the same poles as* p. *Suppose* b *is admissible on* $\bar{\Omega}$, *and* (3.4) *and* (3.5) *hold. Then* $\{\chi_k\}_{k\in I}$ *is a Riesz basis for* H.

Remark. Note that (3.5) is true if $\{\alpha_k-\lambda_k\}_{k\in I}\in l_\infty$, because of (3.32).

Proof. In this proof let

$$j_k\in L_k:=\{j\in I\,|\,|\alpha_k-\lambda_j| \text{ is minimized}\} \tag{3.42}$$

so that $|\alpha_k-\lambda_j|\geqq|\lambda_{j_k}-\lambda_j|-|\lambda_{j_k}-\alpha_k|\geqq m|j-j_k|-|\alpha_k-\lambda_j|$, which implies that

$$|\alpha_k-\lambda_j|\geqq(m/2)|j-j_k|. \tag{3.43}$$

In particular, this implies that $|\alpha_k-\lambda_j|\geqq m/2$ for $j\neq j_k$, so $L_k=\{j_k\}$ if $|\lambda_{j_k}-\alpha_k|<m/2$. Let

$$\kappa:=\{k\in I\,|\,|\lambda_{j_k}-\alpha_k|<m/2\}, \tag{3.44}$$

$$\begin{aligned}&\mathcal{J}_k=I \quad \text{for } k\notin\kappa,\\ &\mathcal{J}_k=I\backslash\{j_k\} \quad \text{for } k\in\kappa;\end{aligned} \tag{3.45}$$

therefore,

$$|\alpha_k-\lambda_j|\geqq m/2 \quad \text{for } j\in\mathcal{J}_k, \quad k\in I. \tag{3.46}$$

Taking Lemma 9 into account, we see that $\{\chi_k\}_{k\in I}$ is a Riesz basis for H if (3.23) is true when y_j is given by (3.25). Therefore, we want a bound on

$$\sum_{k\in I}|y_k|^2=\sum_{k\in I}\left|\sum_{j\in I}x_jq(\lambda_j)b_kp'(\lambda_k)/[b_jp'(\lambda_j)q'(\alpha_k)(\lambda_j-\alpha_k)]\right|^2. \tag{3.47}$$

Taking (3.12) and (3.13) into account, we have that

$$\begin{aligned}&\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jq(\lambda_j)b_kp'(\lambda_k)/[b_jp'(\lambda_j)q'(\alpha_k)(\lambda_j-\alpha_k)]\right|^2\\ &\qquad\leqq M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_k/p'(\lambda_j)(\lambda_j-\alpha_k)\right|^2\end{aligned} \tag{3.48}$$

for some M. Letting a satisfy (3.35), we can write this as

$$\begin{aligned}&M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_k/p'(\lambda_j)(\alpha_j-\lambda_k-a)\right.\\ &\qquad\left.+\sum_{j\in\mathcal{J}_k}x_jb_k(\alpha_j-\lambda_j+\alpha_k-\lambda_k-a)/[p'(\lambda_j)(\alpha_j-\lambda_k-a)(\lambda_j-\alpha_k)]\right|^2\\ &\leqq M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_k/p'(\lambda_j)(\alpha_j-\lambda_k-a)\right|^2\\ &\quad+M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_k(\alpha_j-\lambda_j)/[p'(\lambda_j)(\alpha_j-\lambda_k-a)(\lambda_j-\alpha_k)]\right|^2\\ &\quad+M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_k(\alpha_k-\lambda_k)/[p'(\lambda_j)(\alpha_j-\lambda_k-a)(\lambda_j-\alpha_k)]\right|^2\\ &\quad+M\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k}x_jb_ka/[p'(\lambda_j)(\alpha_j-\lambda_k-a)(\lambda_j-\alpha_k)]\right|^2.\end{aligned}$$

We will refer to these last four sums as $S_1+S_2+S_3+S_4$. By Lemma 8, S_1 is bounded by $M\sum_{j\in I}|x_j|^2$ for some M. Taking (3.13) into account, we have that

$$S_2 \leqq M \sum_{k\in I}\sum_{j\in\mathscr{I}_k}|x_j b_k/(\alpha_j-\lambda_k-a)|^2\left\{\sum_{j\in\mathscr{I}_k}|(\alpha_j-\lambda_j)/(\lambda_j-\alpha_k)|^2\right\}. \tag{3.49}$$

When we use (3.4), the term in brackets is seen to be less than or equal to

$$\sum_{j\in\mathscr{I}_k}|b_j/(\lambda_j-\alpha_k)|^2=\sum_{j\in\mathscr{I}_k}|b_j/(\lambda_j-\alpha_k-\rho)|^2|(\lambda_j-\alpha_k-\rho)/(\lambda_j-\alpha_k)|^2. \tag{3.50}$$

Since the term $|(\lambda_j-\alpha_k-\rho)/(\lambda_j-\alpha_k)|$ is uniformly bounded for $j\in\mathscr{I}_k$ and $k\in I$, from (3.44), the right side of (3.50) is uniformly bounded, from (3.38).

Therefore the term in brackets in (3.49) is uniformly bounded for $j\in\mathscr{I}_k$ and $k\in I$, so

$$S_2 \leqq M_1\sum_{j\in I}|x_j|^2\left\{\sum_{k\in I}|b_k/(\alpha_j-\lambda_k-a)|^2\right\}\leqq M\sum_{j\in I}|x_j|^2$$

for some M and M_1, again because of (3.38).

To estimate S_3, use (3.38) and hypothesis (3.35) to see that

$$\begin{aligned}S_3 &\leqq M\sum_{k\in I}\sum_{j\in\mathscr{I}_k}|x_j b_k/(\alpha_j-\lambda_k-a)|^2\left\{\sum_{j\in\mathscr{I}_k}|(\alpha_k-\lambda_k)/(\lambda_j-\alpha_k)|^2\right\}\\ &\leqq M\sum_{j\in I}|x_j|^2\sum_{k\in I}|b_k/(\alpha_j-\lambda_k-a)|^2\leqq M_1\sum_{j\in I}|x_j|^2\end{aligned}$$

for some M and M_1. Similar estimates are easily seen to hold for S_4, so

$$\sum_{k\in I}\left|\sum_{j\in\mathscr{I}_k}x_j q(\lambda_j)b_k p'(\lambda_k)/[b_j p'(\lambda_j)q'(\alpha_k)(\lambda_j-\alpha_k)]\right|^2\leqq M\sum_{j\in I}|x_j|^2 \tag{3.51}$$

for some M. Comparing (3.51) with (3.47), we see that (3.23) is true if

$$\sum_{k\in\kappa}|x_{j_k}q(\lambda_{j_k})b_k p'(\lambda_k)/[b_{j_k}p'(\lambda_{j_k})q'(\alpha_k)(\lambda_{j_k}-\alpha_k)]|^2\leqq M\sum_{j\in I}|x_j|^2$$

(cf. (3.44)). Using (3.12), (3.13), and (3.17), we see that this is true if

$$\sum_{k\in\kappa}|x_{j_k}b_k/b_{j_k}|^2\leqq M\sum_{j\in I}|x_j|^2. \tag{3.52}$$

LEMMA 12. $\{b_k/b_{j_k}\}_{k\in\kappa}\in l_\infty$.

Proof. To prove Lemma 12, we first need to show that

$$\left\{\sum_{k\in I}|b_k/(\alpha_k-\lambda_j-a)|^2\right\}\in l_\infty.$$

This is done by an argument similar to that used to analyze (3.48). The proof then proceeds as does the proof of Lemma 10, so we will omit the details and consider Lemma 12 proved. The proof of Lemma 11 is then complete, because Lemma 12 implies (3.52). □

We now use this theorem to show that h^* is an admissible feedback element.

Proof of Theorem 2. We will prove this by referring to Proposition 1. $\{\chi_k\}_{k\in I}$ is a Riesz basis for H, by Theorem 11. Lemma 7 shows that (2.8) is true. Suppose x is in the set on the left side of (2.9), so x can be written as $\sum_{k\in I}(x_k/\alpha_k)\chi_k$, where $\{x_k\}_{k\in I}\in l_2$. Using condition (6) in Definition 3 in the Appendix, we see that $\{1/\alpha_k\}_{k\in I}\in l_2$. This implies that $\{x_k/\alpha_k\}_{k\in I}\in l_1$, so $x\in\mathscr{D}(h^*)$ and (2.9) is true. All of the hypotheses of Proposition 1 are satisfied, so h^* is admissible and $A+bh^*$ has eigenvalues at $\{\alpha_k\}_{k\in I}$ and eigenvectors $\{\chi_k\}_{k\in I}$. □

We can get similar results with different restrictions on $\{\alpha_k\}_{k\in I}$. We will go through a similar procedure to prove Theorem 3.

THEOREM 13. *Suppose $\{\alpha_k\}_{k\in I}$ is the zero set of a cardinal function q which has the same poles as p. Suppose b is admissible on $\tilde{\Omega}$, and (3.7) is true. Then $\{\chi_k\}_{k\in I}$ is a Riesz basis for H.*

Proof. As in the proof of Theorem 11, the above is true if (3.23) holds when y_j is given by (3.25). We see that the left side of (3.48) is less than or equal to

$$\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k} x_j q(\lambda_j)b_k p'(\lambda_k)/[b_j p'(\lambda_j)q'(\alpha_k)(\alpha_j-\lambda_k-a)]\right|^2$$
$$+\sum_{k\in I}\left|\sum_{j\in\mathcal{J}_k} M_{j,k}x_j q(\lambda_j)b_k p'(\lambda_k)/[b_j p'(\lambda_j)q'(\alpha_k)(\alpha_j-\lambda_k-a)]\right|^2,$$

where

$$M_{j,k}=[(\alpha_j-\lambda_j)+(\alpha_k-\lambda_k)-a]/(\alpha_k-\lambda_j),$$

which we write as S_1+S_2.

We use (3.12), (3.13), and (3.7) to show that $|q(\lambda_j)p'(\lambda_k)/b_j q'(\alpha_k)|$ is bounded, so Lemma 9 implies that $S_1\leqq M\sum_{j\in I}|x_j|^2$ for some M. To estimate S_2, we need the following result.

LEMMA 14. $\{M_{j,k}\}_{j\in\mathcal{J}_k,k\in I}\in l_\infty$.

Proof of Lemma 14. First note that we have

$$|(\alpha_j-\lambda_j)/(\alpha_k-\lambda_j)|\leqq M|b_j/(\alpha_k-\lambda_j)| \tag{3.53}$$

for some M, using (3.7). Using (3.38) and (3.46), we see that this is uniformly bounded for $j\in\mathcal{J}_k$ and $k\in I$. The term $|a/(\alpha_k-\lambda_j)|$ is also uniformly bounded for $j\in\mathcal{J}_k$ and $k\in I$, by (3.46). Hence, we are interested in showing that the term

$$N_{j,k}:=(\alpha_k-\lambda_k)/(\alpha_k-\lambda_j)$$

is uniformly bounded for $j\in\mathcal{J}_k$ and $k\in I$. Let

$$\eta_k:=(\lambda_k-\alpha_k)/b_k,\qquad \mu_{j,k}:=b_k/(\alpha_j-\lambda_k-a).$$

Formulae (3.7) and (3.38) imply that

$$\{\eta_k\}_{k\in I}\in l_2,\qquad |\mu_{j,k}|\leqq M \tag{3.54}$$

for some M independent of $j\in I$ and $k\in I$. It is easily checked that

$$N_{j,k}=\{[(\alpha_j-\lambda_j)/(\alpha_k-\lambda_j)]-1\}\eta_k\mu_{j,k}\{(\alpha_j-\lambda_k-a)/(\alpha_j-\alpha_k)\}.$$

We can use (3.38), (3.46), and (3.7) to show that $\{[(\alpha_j-\lambda_j)/(\alpha_k-\lambda_j)]-1\}$ is uniformly bounded for $j\in\mathcal{J}_k$ and $k\in I$. We can easily verify that

$$(\alpha_j-\lambda_k-a)/(\alpha_j-\alpha_k)=[\eta_k\mu_{j,k}+(\alpha_k-\lambda_j)/(\alpha_k-\lambda_j-a)]^{-1}.$$

The term $(\alpha_k-\lambda_j)/(\alpha_k-\lambda_j-a)$ is bounded below for $j\in\mathcal{J}_k$ and $k\in I$, and $\eta_k\mu_{j,k}$ goes to zero as k goes to ∞, uniformly in j. Therefore, there exists K such that $|\eta_k\mu_{j,k}+(\alpha_k-\lambda_j)/(\alpha_k-\lambda_j-a)|\geqq c>0$ for all $|k|\geqq K$ and $j\in I$. Hence, $N_{j,k}$ is uniformly bounded for $j\in\mathcal{J}_k$ and $|k|\geqq K$. Because $\{|\lambda_k-\alpha_k|\}_{|k|<K}$ is a bounded set, when it is combined with the definition in (3.44), we see that $N_{j,k}$ is uniformly bounded for $j\in\mathcal{J}_k$ and $|k|<K$. Thus the proof of Lemma 14 is complete. □

To continue analyzing S_2, we note that

$$S_2 \leqq \sum_{k\in I}\sum_{j\in\mathcal{I}_k} |M_{j,k}x_j b_k p'(\lambda_k)/[p'(\lambda_j)q'(\alpha_k)(\alpha_j-\lambda_k-a)]|^2 \sum_{j\in\mathcal{I}_k} |q(\lambda_j)/b_j|^2$$

$$\leqq M \sum_{k\in I}\sum_{j\in\mathcal{I}_k} |x_j b_k/(\alpha_j-\lambda_k-a)|^2 \left\{ \sum_{j\in\mathcal{I}_k} |q(\lambda_j)/b_j|^2 \right\},$$

for some M, where we used (3.13) and Lemma 14 for the last inequality. The term in brackets is bounded because of (3.2) and (3.57), so (3.38) then implies that this is bounded by $M\sum_{j\in I}|x_j|^2$ for some M. This verifies (3.23) when y_j is given by (3.25), and so the proof of Theorem 13 is complete. □

As in the proof of Theorem 2, the Riesz basis property of $\{\chi_k\}_{k\in I}$ is enough to show that h^* is an admissible feedback element. To show that h^* is bounded, we note that $\{h_j\}_{j\in I}$ is biorthogonal to $\{\chi_k\}_{k\in I}$ (cf. (3.9)), so $\{h_j\}_{j\in I}$ is a Riesz basis for H'. Since (3.12) and (3.7) imply that

$$\{p(\alpha_k)/b_k\}_{k\in I} \in l_2,$$

the boundedness of h^* follows from (3.13) and (1.12). This completes the proof of Theorem 3. □

Theorem 4 is less general than Theorems 2 or 3, so we can prove it in a more direct fashion. Before we prove Theorem 4, we will show that it is not a special case of either Theorem 2 or Theorem 3.

LEMMA 15. *Assume that the hypotheses of Theorem* 4 *hold. If* $\sigma>0$ *in* (1.3) b *is admissible on* $(0,\infty)$. *For any* $\sigma\in[0,\pi/2)$, b *is not admissible on* $\bar{\bar{\Omega}}$.

Proof. Referring to Definition 1, we have that the input element b is admissible on $(0,\infty)$ if (3.34) is true for each $t\in(0,\infty)$ for some $M=M(t)$. For each $t\in(0,\infty)$, let $\varphi(\lambda)=\int_0^t e^{-\lambda s}u(s)\,ds$, so the left side of (3.34) is

$$\sum_{k\in I} |b_k e^{\lambda_k t}|^2 |\varphi(\lambda_k)|^2. \tag{3.55}$$

It is well known (see, for instance, [6]) that $\sum_{k\in I}|\varphi(\lambda_k)|^2 \leqq M_1(t)\int_0^t |u(s)|^2\,ds$ for some $M_1(t)$. If $\sigma>0$, then $-\mathrm{Re}\,(\lambda_k)\geqq \tilde{m}k^2$ for some $\tilde{m}>0$ and large enough k, and $|b_k e^{\lambda_k t}|^2$ is bounded. Hence (3.55) is less than or equal to $M(t)\int_0^t |u(s)|^2\,ds$ for some $M(t)$, so (3.34) is verified. Therefore b is admissible on the positive real axis if $\sigma>0$.

To see that b is not admissible on $\bar{\bar{\Omega}}$, let $t=e^{-i\sigma}$. Plugging $u(s)=e^{\lambda_j s}$ into (3.34), we see that if b is admissible on $\bar{\bar{\Omega}}$, then

$$\sum_{k\in I} \left| b_k \int_{[0,t]} e^{\lambda_k(t-s)}\, e^{\lambda_j s}\,ds \right|^2 \leqq M \int_{[0,t]} e^{2\mathrm{Re}(\lambda_j s)}|ds|$$

for some M. Therefore, if b is admissible on $\bar{\bar{\Omega}}$,

$$\left| b_j \int_{[0,t]} e^{\lambda_j(t-s)}\, e^{\lambda_j s}\,ds \right|^2 \leqq M \int_{[0,t]} e^{2\mathrm{Re}(\lambda_j s)}|ds|.$$

When we compute the integrals above, this becomes

$$|b_j \exp(\lambda_j e^{-i\sigma})|^2 \leqq M\{\exp(2\,\mathrm{Re}\,(\lambda_j e^{i\sigma}))=1\}/2\,\mathrm{Re}\,(\lambda_j e^{-i\sigma}) \tag{3.56}$$

(if $\mathrm{Re}\,(\lambda_j e^{-i\sigma})$ is zero, then the term on the right is M). When $j>0$, $\mathrm{Re}\,(\lambda_j e^{-i\sigma})$ is bounded above and below, so that (3.56) is true only if $\{b_j\}_{j>0}$ is bounded, which we do not have in this case. Therefore b is not admissible on $\bar{\bar{\Omega}}$. □

Proof of Theorem 4. Once again, we start by showing that (3.23) is true for y_j given by (3.21). Since (3.12) is true, we see that $\{p(\alpha_k)\}_{k\in I}$ is a bounded sequence. Letting $z_{k,j} = x_k p(\alpha_k) b_j / b_k p'(\lambda_k)(\alpha_j - \lambda_k)$, we have that

$$\sum_{j\in I} |y_j|^2 \leqq 2\left\{ \sum_{j>0} \left| \sum_{k>0} z_{k,j} \right|^2 + \sum_{j>0} \left| \sum_{k<0} z_{k,j} \right|^2 + \sum_{j<0} \left| \sum_{k>0} z_{k,j} \right|^2 + \sum_{j<0} \left| \sum_{k<0} z_{k,j} \right|^2 \right\}. \tag{3.57}$$

We will estimate the first of these sums.

Let $z_k = \alpha_k - \lambda_k$. The hypotheses of Theorem 4 imply that there exist c_2, δ, and $c_1 > 0$ such that $|\lambda_k - \lambda_j| \geqq c_1 |k^2 - j^2|$ for $k > 0$, $j > 0$, and $|k-j| > \delta$, $|z_k| \leqq c_2 k$. Choose δ large enough so that $c = c_1\delta - c_2 > 0$. Then, for $k > 0$, $j > 0$, and $|k-j| > \delta$, $|\alpha_k - \lambda_j| = |\lambda_k - \lambda_j + z_k| \geqq c_1|k^2 - j^2| - c_2 k = c_1|k-j||k+j| - c_2 k \geqq (c_1\delta - c_2)k = ck \geqq M\|z_k\|$ (cf. (3.8)). This implies that for $k > 0$, $j > 0$, and $|k-j| > \delta$, we have that

$$|(\lambda_k - \lambda_j)/(\alpha_k - \lambda_j)| = |1 - z_k/(\alpha_k - \lambda_j)| \leqq M$$

for some M. This implies that

$$|1/(\alpha_k - \lambda_j)| \leqq M/(\lambda_\kappa - \lambda_j) \leqq (M/c_1)/(k^2 - j^2). \tag{3.58}$$

Let $\xi_k = x_k / p'(\lambda_k) a_k$, so that (using the hypotheses of Theorem 4)

$$\begin{aligned} \sum_{j>0} \left| \sum_{k>0} z_{k,j} \right|^2 &\leqq M \sum_{j>0} \left| \sum_{k>0} \xi_k p(\alpha_k) \beta_j j / \beta_k k(\alpha_j - \lambda_k) \right|^2 \\ &\leqq 2M \left\{ \sum_{j>0} \left| \sum_{\substack{k>0 \\ |k-j|>\delta}} \xi_k p(\alpha_k) \beta_j j / \beta_k k(\alpha_k - \lambda_j) \right|^2 \right. \\ &\qquad \left. + \sum_{j>0} \left| \sum_{\substack{k>0 \\ |k-j|\leqq\delta}} \xi_k p(\alpha_k) \beta_j j / \beta_k k(\alpha_k - \lambda_j) \right|^2 \right\} \end{aligned}$$

for some M, which we write as $S_1 + S_2$. Since $\{p(\alpha_k)\}_{k\in I}$ and $\{\beta_j/\beta_k\}_{j,k\in I}$ are bounded and (3.58) is true for $j > 0$, $k > 0$, and $|k-j| > \delta$, there exist M and M_1 such that

$$\begin{aligned} S_1 &\leqq M \sum_{j>0} \left| \sum_{\substack{k>0 \\ |k-j|>\delta}} |j/k(\alpha_k - \lambda_j)| \right|^2 \sum_{k>0} |\xi_k|^2 \\ &\leqq \left\{ M_1 \sum_{k>0} |\xi_k|^2 \right\} \sum_{k>0} (1/k^2) \left[\sum_{\substack{j>0 \\ |k-j|>\delta}} |j/(k+j)(k-j)|^2 \right]. \end{aligned}$$

Since the term in square brackets is easily seen to be bounded independently of $k > 0$, we have that $S_1 \leqq M \sum_{k>0} |\xi_k|^2$ for some M, which is seen to be less than $M \sum_{k>0} |x_k|^2$ for some M, from (3.12).

To show that S_2 is bounded, note that (3.12) implies that

$$S_2 \leqq M \sum_{j>0} \left| \sum_{\substack{k>0 \\ |k-j|\leqq\delta}} \xi_k j/k \right|^2 \leqq M_1 \sum_{j>0} \left\{ \sum_{\substack{k>0 \\ |k-j|\leqq\delta}} (j/k)^2 \right\} \sum_{\substack{k>0 \\ |k-j|\leqq\delta}} |\xi_k|^2$$

for some M and M_1. The term in brackets is easily seen to be bounded for any j, so that

$$S_2 \leqq M \sum_{k>0} |\xi_k|^2 \sum_{\substack{|k-j|\leqq\delta \\ k>0}} 1 \leqq M_1 \sum_{k>0} |x_k|^2$$

for some M and M_1. This shows that the first sum on the right side of (3.57) is bounded by $M \sum_{k>0} |x_k|^2$ for some M. We handle the other sums in (3.57) in a similar way, showing that (3.23) is true for $\{y_j\}_{j\in I}$ given by (3.21).

In order to show that $\{\chi_k\}_{k\in I}$ is a Riesz basis, it suffices to show that (3.23) is true for y_j given by (3.25). This is done by the methods used above. The rest of the theorem is a consequence of Lemma 7 and Proposition 1. □

4. Examples. Many vibrating systems can be written in the form (1.1) considered in this paper. Combining Theorems 2, 3, and 4 with the examples of cardinal functions given in [8], we can get specific eigenvalue specification results for these systems. We now give two examples involving an elastic beam with both ends hinged, and apply the results above to give eigenvalue specification results which, to the knowledge of the author, are new. In the first example the control is a point actuator, which yields an input element b that is admissible on $\tilde{\tilde{\Omega}}$, so we can apply Theorems 2 and 3. In this example the input is not bounded, and the input coefficients do not satisfy (1.7), which are the two typical restrictions on the input (cf. [7], [9], [12]). In the second example the control is a moment force on one end, which yields an input element that is not bounded or even admissible on $\tilde{\tilde{\Omega}}$. However, we can apply Theorem 4 to this example. To the author's knowledge, this is the first eigenvalue specification result for such a system.

Example 1. To describe a structurally damped beam with both ends hinged, we use the Euler-Bernoulli beam model, with the damping term arising from a lateral force on the beam negatively proportional to the rate of bending [10]. Let $w(x, t)$ be the lateral deflection of the beam, let w_t be the derivative of w with respect to time, and let w_x be the derivative with respect to the position $x\in[0, L]$. Assume that the flexural rigidity EI and the density ρ are constant. If there is no external force the beam is modeled by

$$w_{tt}(x, t)-2\gamma(EI/\rho)^{1/2}w_{txx}(x, t)+(EI/\rho)w_{xxxx}(x, t)=0, \qquad 0\leqq\gamma<1, \tag{4.1}$$

$$w(0, t)=0, \quad w(1, t)=0, \quad w_{xx}(0, t)=0, \quad w_{xx}(1, t)=0, \tag{4.2}$$

with initial conditions

$$w(x, 0)=w_0, \qquad w_t(x, 0)=w_1. \tag{4.3}$$

If we have a point actuator at $x_0\in[0, L]$, the zero on the right-hand side of (4.1) is replaced by $\delta(x-x_0)u(t)$, where u is a scalar control force and δ is the Dirac delta function. Without loss of generality, let $L=1$.

To put this in state space form, let $X=L^2[0, 1]$, and

$$B: X\to X: w\to(EI/\rho)w_{xxxx},$$

$$\mathcal{D}(B)=\{w\in H^4[0, 1]\,|\,w(0)=w(1)=w_{xx}(0)=w_{xx}(1)=0\}.$$

Then B has a set of eigenvectors $\{\Phi_k\}_{k\in Z^+}$ which form an orthonormal basis for X, and associated eigenvalues $\{\sigma_k\}_{k\in Z^+}$, where $\Phi_k(x)=2^{1/2}\sin(\pi kx)$ and $\sigma_k=(EI/\rho)(\pi k)^4$. With this choice of $\mathcal{D}(B)$, $B^{1/2}$ is given by the differential operator $B^{1/2}w=(EI/\rho)^{1/2}w_{xx}$ and

$$\mathcal{D}(B^{1/2})=\{w\in H^2[0, 1]\,|\,w(0)=w(1)=0\}.$$

Equations (4.1)-(4.3) then become (now we use a dot to denote differentiation with respect to time)

$$\ddot{w}+2\gamma B^{1/2}\dot{w}+Bw=\hat{b}u, \qquad 0<\gamma<1, \tag{4.4}$$

where $\hat{b}=\delta(x-x_0)$. If the control is distributed over the spatial extent of the beam, $\hat{b}$ is replaced by some element of X. Formula (4.4) with a homogeneous left side is a general model for structural damping, given by Chen and Russell in [1], when the

undamped, uncontrolled system is given by $\ddot{w} + Bw = 0$. Unfortunately, it is not always that easy to identify the square root of the operator B. For instance, in the cantilever case discussed in [10], $B^{1/2}$ is a pseudodifferential operator, so Russell introduces a damping term that has a more direct physical interpretation. However, numerical tests in [8] and [13] indicate that in the cantilever case, the damping term $-2\gamma(EI/\rho)^{1/2}w_{txx}(x, t)$ still leads to a system which has eigenvalues between two curves of the form Γ_a when γ is small.

Let

$$Y := \mathcal{D}(B^{1/2}), \qquad H := Y \oplus X,$$

and

$$A = \begin{bmatrix} 0 & I \\ -B & -2\gamma B^{1/2} \end{bmatrix}, \tag{4.5}$$

with domain $\mathcal{D}(A) = \mathcal{D}(B) \oplus Y \subset H$. Let

$$z(t) = [w(t), \dot{w}(t)]^T, \quad z_0 = [w_0, w_1]^T, \quad b = [0, \delta(\cdot - x_0)]^T.$$

Therefore (4.1) and (4.2) with the point actuator control at x_0 can be written in the form (1.1). To see that A generates a holomorphic semigroup on H it is sufficient to note that A has eigenvectors

$$\varphi_{\pm k} := \frac{1}{\sqrt{2}} \begin{bmatrix} \sigma_k^{-1/2}\Phi_k \\ e^{\pm i\eta}\Phi_k \end{bmatrix}, \quad e^{\pm i\eta} = (-\gamma \pm \sqrt{1-\gamma^2}), \quad k \in Z^+ \tag{4.6}$$

with associated eigenvalues

$$\lambda_{\pm k} = \sigma_k^{1/2} e^{\pm i\eta}, \tag{4.7}$$

that these eigenvectors are easily seen to form a Riesz basis for H, and that the eigenvalues lie on Γ_0 (cf. (1.3) with $\sigma = \eta - \pi/2$. Let I be the index set $Z^+ \cup Z^-$, so that A generates the semigroup given by (1.5). For our purposes, it is convenient to think of the dual space of H as being

$$H' := Y' \oplus X,$$

where

$$Y' := \mathcal{D}(B^{1/2})' = \left\{ \sum_{k \in I} x_k \Phi_k \Big| \sum_{k \in I} |x_k^2/\sigma_k| < \infty \right\} = \mathcal{D}(B^{-1/2}).$$

The inner product for $[z_1, z_2]^T \in H'$ and $[y_1, y_2]^T \in H$ is

$$\langle [z_1, z_2]^T, [y_1, y_2]^T \rangle := \int_0^1 \left(B^{-1/2} z_1(x) \overline{B^{1/2} y_1(x)} + z_2(x) \overline{y_2(x)} \right) dx. \tag{4.8}$$

The domain of A^*, the adjoint of A, is the subset of H' given by $\mathcal{D}(A^*) = X \oplus Y$. It is easy to see that

$$\psi_{\pm k} := (1/\sqrt{2} \sin(\eta)) \begin{bmatrix} -(\pm i\sigma_k^{1/2} e^{\pm i\eta})\Phi_k \\ \pm i\Phi_k \end{bmatrix}, \qquad k \in Z^+ \tag{4.9}$$

is an eigenvector of A^* with associated eigenvalue λ_k, that $\{\psi_k\}_{k \in I}$ is a Riesz basis for H', and that $\langle \varphi_j, \psi_k \rangle = \delta_{j,k}$, so that $\{\psi_k\}_{k \in I}$ is the dual basis to $\{\varphi_k\}_{k \in I}$.

The input coefficients can be easily computed in this case: they are

$$b_{\pm k} = \langle b, \psi_{\pm k} \rangle = \langle \delta(\cdot - x_0), \pm i\Phi_k \rangle = \pm i\Phi_k(x_0) = \pm i \sin(\pi k x_0)/\sin(\eta).$$

There are no nonzero values of x_0 for which $\{b_k\}_{k \in I}$ is in l_2. Since $\{b_k\}_{k \in I}$ is bounded,

it is easy to use methods in [5] to show that b is admissible on $\tilde{\tilde{\Omega}}$, so we can apply Theorems 2 and 3 to this system.

There is no value of x_0 for which b satisfies (1.7). If x_0 is rational, b_k is zero infinitely many times, and the $\{b_k\}_{k\in J}$ (cf. (1.8)) satisfies (1.7), so we can apply the methods in [7] to the system. (In fact, the results in [7] will be better in this case, since the only restriction on $\{\alpha_k\}_{k\in I}$ is that it is the zero set of a cardinal function with the same poles as p.) If x_0 is irrational, $b_k \neq 0$ for any $k \in I$, but (1.7) is not satisfied, and the author knows of no other eigenvalue specification results for this system. The following corollary of Theorems 2 and 3 gives us a class of eigenvalues that can be realized for the closed loop system.

COROLLARY 16. *Suppose* $\alpha_k = \lambda_k + \eta_k \sin(\pi k x_0)$, *where* $\{\eta_k\}_{k\in I} \in l_\infty$. *Then* $\{\chi_k\}_{k\in I}$ *given by* (1.10) *is a Riesz basis for* H, h^* *given by* (1.12) *with domain* (3.6) *is an admissible feedback element, and* $A+bh^*$ *has eigenvectors* $\{\chi_k\}_{k\in I}$ *and eigenvalues* $\{\alpha_k\}_{k\in I}$. *If* $\{\eta_k\}_{k\in I} \in l_2$, *then* h^* *is bounded, i.e.,* $h^* \in H$.

Proof. It is clear that (3.4) is satisfied. Since $\{\sin(\pi k x_0)\}_{k\in I}$ is bounded, $\{\alpha_k - \lambda_k\}_{k\in I}$ is bounded, so (3.5) is also satisfied (see the remark after the statement of Theorem 11). Let

$$p(\lambda) = \prod_{k\in I} (\lambda - \lambda_k)/(\lambda - \lambda_k - m), \quad q(\lambda) = \prod_{k\in I} (\lambda - \alpha_k)/(\lambda - \lambda_k - m), \tag{4.10}$$

where m is chosen so that there is a constant c such that $\{\alpha_k\}_{k\in I}$ lies to the right of Γ_c and $\{\lambda_k - m\}_{k\in I}$ lies to the left of Γ_c. It is shown in [8] that p and q are cardinal functions. Therefore we can apply Theorem 2 to this system if $\{\eta_k\}_{k\in I} \in l_\infty$. If $\{\eta_k\}_{k\in I} \in l_2$ we can apply Theorem 3. □

To illustrate this, let $x_0 = 1/\pi$, and the damping factor $\gamma = .2588$, and we will determine the feedback law that realizes the closed-loop eigenvalues $\alpha_{\pm k} = \lambda_{\pm k} - 5\sin(k)$. Let p and q be given by (4.10), with $m = 6$. We will approximate p and q by finite products to estimate $p(\alpha_j)$, $q(\lambda_k)$, $p'(\lambda_k)$, and $q'(\alpha_j)$ so that the last two approximations p_n and p_{n+1} satisfy $|(p_n/p_{n+1}) - 1| < .001$. Using formulae (1.11) and (1.12), we can find h^* in the form $\sum_{j>0} c_j\psi_j + \overline{c_j}\psi_{-j}$, where the first nine coefficients are

$$\{c_j\}_{j=1}^9 = \{.8804 + 5.1598i,\ 1.3474 + 4.4128i,\ 1.2395 + 4.3680i,\ .6952 + 4.6051i,$$
$$.0532 + 4.8249i,\ -.2895 + 4.9092i,\ -.2075 + 4.8877i,\ .0808 + 4.8120i,$$
$$.2614 + 4.7571i\}.$$

In this case h^* is not a bounded element, but we can truncate the series for h^* to get a bounded element. In [8] the effect of the truncated feedback element on the distributed parameter system is studied.

If h^* is written as $[h_1(s), h_2(s)]$, then the control is given by

$$u(t) = \langle h^*, z(t)\rangle = (\text{cf. } (4.8)) \int_0^1 [\tilde{h}_1(s)\overline{w_{xx}(x,t)} + h_2(x)\overline{w_t(x,t)}]\,dx, \tag{4.11}$$

where

$$\tilde{h}_1(x) = B^{-1/2}h_1(x) = \sum_{k>0} f_k\sqrt{2}\sin(\pi k x),$$
$$h_2(x) = \sum_{k>0} g_k\sqrt{2}\sin(\pi k x), \tag{4.12}$$

with the first nine coefficients given by

$$\{f_k\}_{k=1}^9 = \{-.7101, .2333, .0977, -.7619, -1.7531, -2.2697, -2.1456, -1.7091, -1.4329\},$$
$$\{g_k\}_{k=1}^9 = \{-7.5545, -6.4608, -6.3952, -6.7423, -7.0642, -7.1875, -7.1561,$$
$$-7.0452, -6.9649\}.$$

Example 2. We now consider a beam with both ends hinged and a moment force $u(t)$ at one end. This is modeled by (4.1), where we derive the correct boundary conditions for the damped beam by computing the time rate of change of the energy associated with (4.1). The energy for the Euler-Bernoulli beam is

$$(\mathscr{E}(w))(t) := \frac{1}{2}\int_0^1 (EI|w_{xx}(x,t)|^2 + \rho|w_t(x,t)|^2)\,dx. \tag{4.13}$$

Integration by parts twice yields

$$\frac{d(\mathscr{E}(w)(t))}{dt} + (EI\rho)^{1/2}\int_0^1 |w_{xt}(x,t)|^2\,dx$$
$$= \int_0^1 w_t(x,t)\{(EI)w_{xxxx}(x,t) + \rho w_{tt}(x,t) - 2\gamma(\rho EI)^{1/2}w_{txx}(x,t)\}\,dx$$
$$+ \{EIw_{xx}(x,t)\}w_{tx}(x,t) + \{2\gamma(\rho EI)^{1/2}w_x(x,t) - EIw_{xxx}(x,t)\}w_t(x,t)\Big|_0^1.$$

If w satisfies the differential equation (4.1) and the boundary conditions

$$w(0,t)=0, \quad w(1,t)=0, \quad w_{xx}(0,t)=0, \quad w_{xx}(1,t)=u(t), \tag{4.14}$$

then

$$\frac{d(\mathscr{E}(w)(t))}{dt} = EIu(t)w_{tx}(1,t) - (EI\rho)^{1/2}\int_0^1 |w_{tx}(x,t)|^2\,dx,$$

which is the correct form for the rate of change of the energy [10].

In this case (4.1) and (4.14) can also be written as (4.4), where the input element $\hat{b}$ will be determined from the boundary conditions (4.14). We will follow the procedure in [5]. We need two extensions of A. One extension is $\hat{A}$: $H \to \mathscr{H}$, given by (2.3). In the cases under consideration,

$$\hat{A}\left(\sum_{k\in I} x_k\varphi_k\right) = \sum_{k\in I} x_k\lambda_k\varphi_k$$

and

$$\mathscr{H} = \left\{\left(\sum_{k\in I} x_k\varphi_k\right)\,\middle|\, \sum_{k\in I}|x_k/\lambda_k|^2 < \infty\right\}.$$

To get the other extension, we first define an extension of B. Let

$$J: X \to X: w \to (EI/\rho)w_{xxxx},$$
$$\mathscr{D}(J) = \{w \in H^4[0,1] \mid w(0) = w(1) = w_{xx}(0) = 0\}.$$

If $\hat{w}(x) := (x^3 - x)/6$, then $\hat{w}$ is in $\mathscr{D}(J)$, and $\hat{w}_{xx}(1) = 1$. We can therefore write any element of $\mathscr{D}(J)$ as $\xi + u\hat{w}$, where $\xi \in \mathscr{D}(B)$ and u is a scalar. Let

$$L: H \to H: Lz = \begin{bmatrix} 0 & I \\ -J & -2\gamma B^{1/2} \end{bmatrix} z, \qquad \mathscr{D}(L) = \mathscr{D}(J) \oplus \mathscr{D}(B^{1/2}), \tag{4.15}$$

so that L is an extension of A. We can write $z \in \mathscr{D}(L)$ as $\xi + u[\hat{w}, 0]^T$, with u a scalar and $\xi \in \mathscr{D}(A)$. We would like the solution $z(t) = [w(t), \dot{w}(t)]^T$ to be in $\mathscr{D}(L)$. In fact, we can write the control system (4.1), (4.14) as

$$\dot{z}(t) = Lz(t), \quad z(t) = \xi(t) + u(t)\begin{bmatrix} \hat{w} \\ 0 \end{bmatrix}, \quad \xi(t) \in \mathscr{D}(A). \tag{4.16}$$

We will rewrite this in the form $\dot{z}(t) = \hat{A}z(t) + (L - \hat{A})z(t)$, where the term $(L - \hat{A})z(t)$ turns out to be of the form $bu(t)$.

Since $L\xi = \hat{A}\xi$, $Lz - \hat{A}z = uL[\hat{w}, 0]^T - u\hat{A}[\hat{w}, 0]^T = -u\hat{A}[\hat{w}, 0]^T$. To compute $\hat{A}[\hat{w}, 0]^T$, let $[\eta_1, \eta_2]^T \in \mathcal{D}(A^*) = X \oplus Y$ (note that this implies that $\eta_2(0) = \eta_2(1) = 0$); therefore,

$$\begin{aligned}\langle [\eta_1, \eta_2]^T, \hat{A}[\hat{w}, 0]^T \rangle &= \langle A^*[\eta_1, \eta_2]^T, [\hat{w}, 0]^T \rangle \\ &= \langle [-B\eta_2, \eta_1 - 2\gamma B^{1/2}\eta_2]^T, [\hat{w}, 0]^T \rangle \\ &= -\langle B^{-1/2}B\eta_2, B^{1/2}\hat{w} \rangle_{L^2[0,1]} \\ &= (-EI/\rho) \int_0^1 \overline{\hat{w}_{xx}(x)} \eta_{2xx}(x)\, dx \\ &= (\text{integrating by parts twice}) = (-EI/\rho)\eta_{2x}(1).\end{aligned}$$

Hence $-\hat{A}[\hat{w}, 0]^T = (EI/\rho)[0, \delta'(\cdot - 1)]^T := b$, and (4.15) becomes

$$\dot{z}(t) = \hat{A}z(t) + bu(t),$$

which is an equation in $\mathcal{H}$. The expansion coefficients of b are $b_{\pm k} = \langle b, \psi_{\pm k} \rangle = \pm i 2^{-1/2}(EI/\rho)\Phi_k'(1)/\sin(\eta) = \pm i(EI/\rho)\pi k(-1)^k/\sin(\eta)$ for $k \in Z^+$, since $\Phi_k(x) = 2^{1/2}\sin(\pi k x)$. Hence we can represent the input element in the form (1.2), and this system satisfies the hypotheses of Theorem 4. Therefore, we can realize any closed-loop eigenvalues such that there is a cardinal function q with the same poles as p, and $\{(\alpha_k - \lambda_k)/b_k\}_{k \in I} \in l_\infty$. For instance, if $\alpha_k = \lambda_k + \eta_k$, where $\eta_k \in l_\infty$, p and q given by (4.10) are shown to be cardinal functions in [8], so $A + bh^*$ has eigenvalues at $\{\alpha_k\}_{k \in I}$ if h^* is given by (4.12).

As an example we consider (4.1) and (4.14) with the damping factor $\gamma = .2588$. We can realize $\alpha_k = \lambda_k - 2$ by the control law (4.11), where $\tilde{h}_1$ and h_2 are given by (4.12), with the first nine coefficients given by

$$\{f_k\}_{k=1}^9 = \{.1790, -.0733, .0540, -.0434, .0279, -.0159, .0160, -.0152, .0141\},$$

$$\{g_k\}_{k=1}^9 = \{.9578, -.4535, .2995, -.2240, .1754, -.1419, .1228, -.1079, .0961\}.$$

Appendix: Definition of cardinal function and consequences.

DEFINITION 3. Let p be a meromorphic function with zero set $\{\lambda_k\}_{k \in I}$ and pole set $\{\mu_k\}_{k \in I}$. Assume that all λ_k's and μ_k's are distinct, and that $I = Z^+ \cup Z^-$. Then p is a cardinal function if:

(1) There exist real constants d, c, and m_1, with $d < c$, such that $|p(\lambda)| \leqq m_1$ for all $\lambda \in \{\lambda > \Gamma_c\} \cup \{\lambda < \Gamma_d\}$.

(2) There exist real constants a, b, and m_2 with $b < a$ such that $|p(\lambda)| \geqq m_2 > 0$ for all $\lambda \in \{\lambda > \Gamma_a\} \cup \{\lambda < \Gamma_b\}$.

(3) There exists $m_3 > 0$ such that $|p'(\lambda_k)|$, $|p'(\mu_k)/(p(\mu_k))^2| \geqq m_3$ for all $k \in I$.

(4) There exist paths $\{\Lambda_j\}_{j \in Z}$ between Γ_a and Γ_b such that
   (a) there exists m_4, independent of j, such that $|p(\lambda)| \geqq m_4 > 0$ for $\lambda \in \Lambda_j$;
   (b) $\{\text{length}(\Lambda_j)\}_{j \in Z}$ is bounded;
   (c) $(\inf_{\lambda \in \Lambda_j} \operatorname{Im}(\lambda)) \longrightarrow_{j \to \infty} \infty$ and $(\sup_{\lambda \in \Lambda_j} \operatorname{Im}(\lambda)) \longrightarrow_{j \to -\infty} -\infty$.

(5) There exist paths $\{\tilde{\Lambda}_j\}_{j \in Z}$ between Γ_c and Γ_d such that
   (a) there exists $m_5 > 0$, independent of j, such that $|p(\lambda)| \leqq m_5$ for $\lambda \in \tilde{\Lambda}_j$;
   (b) $\{\text{length}(\tilde{\Lambda}_j)\}_{j \in Z}$ is bounded;
   (c) $(\inf_{\lambda \in \tilde{\Lambda}_j} \operatorname{Im}(\lambda)) \longrightarrow_{j \to \infty} \infty$ and $(\sup_{\lambda \in \tilde{\Lambda}_j} \operatorname{Im}(\lambda)) \longrightarrow_{j \to -\infty} -\infty$.

(6) $\inf_{k>0}\{\operatorname{Im}(e^{-i\sigma}(\lambda_k - \lambda_{k-1})), \operatorname{Im}(e^{i\sigma}(\lambda_{-k} - \lambda_{-k+1}))\} > 0$, $\operatorname{Im}(\lambda_k) \geqq 0$ for $k > 0$, and $\operatorname{Im}(\lambda_k) \leqq 0$ for $k < 0$.

(7) $\inf_{k>0}\{\operatorname{Im}(e^{-i\sigma}(\mu_k-\mu_{k-1})), \operatorname{Im}(e^{i\sigma}(\mu_{-k}-\mu_{-k+1}))\}>0$, $\operatorname{Im}(\mu_k)\geqq 0$ for $k>0$, and $\operatorname{Im}(\mu_k)\leqq 0$ for $k<0$.

Remark 1. Two classes of cardinal functions are constructed in [8]. In one class, the zeros and poles are spaced asymptotically linearly, and in the other class they grow faster than linearly.

Remark 2. The definition of cardinal function given in [6], [7] requires that

$$|p'(\lambda_k)|\leqq M \quad \text{and} \quad |p'(\mu_k)/(p(\mu_k))^2|\leqq M$$

for some M independent of k. We also need this for results in § 3. To show that this is a consequence of that definition, let C_k be a small circle oriented counterclockwise with center λ_k and radius ε, where ε is chosen so that $p(\lambda)$ is bounded on all C_k. Then

$$|p'(\lambda_k)| = \left|\left(\frac{1}{2}\pi i\right)\int_{C_k}[p(\zeta)/(\zeta-\lambda_k)^2]\,d\zeta\right|,$$

which is easily seen to be bounded independent of k. We could do the same with p replaced by $1/p$ to show that $|p'(\mu_k)/(p(\mu_k))^2|$ is bounded independent of k.

To verify (3.12), let C be the curve $\Gamma_\alpha \cup \Gamma_\beta$, where $c<\alpha<b<a<\beta$, Γ_α is oriented from top to bottom, and Γ_β is oriented from bottom to top. When we use Cauchy's Theorem, we can easily show that

$$\begin{aligned}|p(\alpha_k)-p(\lambda_j)| &= (1/2\pi)\left|\int_C[p(\zeta)/(\alpha_k-\zeta)]\,d\zeta-\int_C[p(\zeta)/(\lambda_j-\zeta)]\,d\zeta\right| \\ &\leqq (1/2\pi)|\alpha_k-\lambda_j|\left\{\int_C|p(\zeta)/(\lambda_j-\zeta)(\alpha_k-\zeta)||d\zeta|\right\}.\end{aligned}$$

Since p is bounded on C, and λ_j and α_k lie between Γ_a and Γ_b, it is easy to see that the term in brackets is uniformly bounded. Since $p(\lambda_j)=0$, (3.12) follows.

REFERENCES

[1] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 33 (1982), pp. 433-454.

[2] B. M. N. CLARKE AND D. WILLIAMSON, *Control canonical forms and eigenvalue assignment by feedback for a class of linear hyperbolic systems*, SIAM J. Control Optim., 19 (1981), pp. 711-729.

[3] R. CURTAIN AND A. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes on Control and Inform. Sci. 8, Springer-Verlag, New York, 1978.

[4] L. F. HO, *Spectral assignability of systems with scalar control and application to a degenerate hyperbolic system*, SIAM J. Control Optim., 24 (1986), pp. 1212-1231.

[5] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614-639.

[6] R. REBARBER, *A Laplace transform relevant to holomorphic semigroups*, Proc. Roy. Soc. Edinburgh Sect. A, 105 (1987), pp. 243-258.

[7] ———, *Canonical forms for a class of distributed parameter control systems*, SIAM J. Control Optim., 26 (1988), pp. 1362-1387.

[8] ———, *A class of meromorphic functions used for spectral determination*, to appear.

[9] D. L. RUSSELL, *Closed loop eigenvalue specification for infinite dimensional systems: Augmented and deficient hyperbolic cases*, Tech. Summary Report 2021, Mathematics Research Center, University of Wisconsin, Madison, WI, August 1979.

[10] ———, *Mathematical models for the elastic beam and their control-theoretic implications*, Semigroup Theory and Applications, H. Brezis, M. G. Crandall, and F. Kappel, eds., Longman, New York, 1985.

[11] S.-H. SUN, *On spectrum distribution of completely controllable linear systems*, SIAM J. Control Optim., 19 (1981), pp. 730-743 (translated by L. F. Ho).

[12] R. TEGLAS, *On the control canonical structure of a class of scalar input systems*, SIAM J. Control Optim., 22 (1984), pp. 552-569.

[13] S.-F. C. YU, *Computing the eigenvalues for a damped Euler-Bernoulli beam equation with the Chebyshev spectral method*, Master's thesis, Pennsylvania State University, State College, PA, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 170-198, January 1989

010

CONTINGENT CONES TO REACHABLE SETS OF CONTROL SYSTEMS*

HALINA FRANKOWSKA†

Abstract. High-order necessary conditions for optimality for an optimal control problem are studied via properties of contingent cones to reachable sets along the optimal trajectory. It is shown that the adjoint vector of Pontryagin's maximum principle is normal to the set of variations of reachable sets. Results are applied to study optimal control problems for dynamical systems described by: (1) closed-loop control systems; (2) nonlinear implicit systems; (3) differential inclusions; (4) control systems with jumps.

Key words. high-order maximum principles, implicit control system, closed-loop control system, discontinuous trajectory, reachable set, contingent cone

AMS(MOS) subject classifications. 49B10, 49B34, 49B36, 49E15, 93C15

1. Introduction.

Consider the following optimal control problem in $\mathbb{R}^n$:

$$\text{minimize } g(x(T)) \tag{1.1}$$

over the solutions to the control system

$$x'(t) = f(x(t), u(t)) \quad \text{a.e. in } [0, T], \tag{1.2}$$

$$u(t) \in U \quad \text{is a measurable selection,} \tag{1.3}$$

$$x(0) \in C. \tag{1.4}$$

Let $R(t, C)$ denote its reachable set at time t from the set of initial conditions $C \subset \mathbb{R}^n$ and $T_{R(t,C)}(x_0)$ the contingent cone to $R(t, C)$ at $x_0 \in \mathbb{R}^n$.

If a trajectory z of the control system (1.2)–(1.4) solves the above problem, then the derivative $g'(z(T))$ is nonnegative in every tangent direction $w \in T_{R(T,C)}(z(T))$, i.e., $g'(z(T))$ belongs to the positive polar cone $T_{R(T,C)}(z(T))^+$ of $T_{R(T,C)}(z(T))$. This is the so-called Fermat rule. Thus we obtain the necessary conditions that allow us to test whether a given trajectory z is optimal whenever we can characterize this positive polar cone. In this paper we study some necessary conditions that can be derived from the above Fermat rule. In the case of nonlinear system, the best we can hope for is to characterize explicitly subsets Q of the tangent cone $T_{R(T,C)}(z(T))$, using variations of the solution $z(\cdot)$.

Then, by duality, $g'(z(T)) \in T_{R(T,C)}(z(T))^+ \subset Q^+$ and the inclusion $g'(z(T)) \in Q^+$ is a necessary condition of optimality. Q is the larger set and Q^+ is the smaller set, so that the necessary condition becomes stronger.

In particular, we prove that the reachable set at time T, $R^L(T)$, of the following linear control system:

$$\begin{cases} w'(t) = \dfrac{\partial f}{\partial x}(z(t), \bar{u}(t))w(t) + v(t) \quad \text{a.e.,} \\ v(t) \in T_{\text{co } f(z(t),U)}(z'(t)), \\ w(0) \in T_C(z(0)) \end{cases} \tag{1.5}$$

(where $\bar{u}$ is a control corresponding to z) is contained in $T_{R(T,C)}(z(T))$. Hence whenever z is optimal, $g'(z(T)) \in R^L(T)^+$.

* Received by the editors April 13, 1987; accepted for publication (in revised form) May 3, 1988.

† Centre de Recherche de Mathématique de la Décision, Université Paris-Dauphine, 75775 Paris, Cedex 16; International Institute for Applied System Analysis, Laxenburg, Austria.

Such inclusion implies easily the celebrated Pontryagin's maximum principle: the solution q of the adjoint system

$$-q'(t) = \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* q(t) \quad \text{a.e. in } [0, T], \tag{1.6}$$

$$q(T) = g'(z(T)) \tag{1.7}$$

satisfies the minimum principle

$$\langle q(t), z'(t)\rangle = \min_{u \in U} \langle q(t), f(z(t), u)\rangle \quad \text{a.e. in } [0, T] \tag{1.8}$$

and the transversality condition

$$q(0) \in T_C(z(0))^+. \tag{1.9}$$

The aim of this paper is to go beyond the maximum principle and to provide some additional properties of the adjoint vector $q(\cdot)$ that can help to eliminate more candidates for optimality than the maximum principle. Let us describe briefly the main ideas.

We introduce the "variations" $\{W(t, z): t \in [0, T]\}$ of $z(\cdot)$, defined by

$$W(t, z) := \{v : \exists h_i \to 0,\ h_i \geqq 0,\ \mu_i \to 0+ \text{such that } z(t+h_i) + \mu_i v \in R(t+h_i, C) + o(\mu_i)\}$$

(in particular, $T_{R(t,C)}(z(t)) \subset W(t, z)$).

For all $0 \leqq t \leqq t+h \leqq T$, define the reachable map $r(h, t): \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ of (1.5) by

$$r(h, t)\xi = \{w(t+h):\ w \in W^{1,1}(t, t+h) \text{ is a solution of (1.5)}, w(t) = \xi\}.$$

We shall prove that for all $t \in [0, T[$, $r(T-t, t)$ maps $W(t, z)$ into $T_{R(T,C)}(z(T))$ and, in particular,

$$r(T, 0) T_C(z(0)) \subset T_{R(T,C)}(z(T)).$$

Thus for all $t \in [0, T]$, $g'(z(T)) \in (r(T-t, t) W(t, z))^+$. If $r(T-t, t)$ is a linear operator, we deduce from the bipolar theorem that $g'(z(T)) \in r(T-t, t)^{*-1}(W(t, z)^+)$, where $r(T-t, t)^*$ is the transpose of $r(T-t, t)$. But the reachable map $r(T-t, t)$ is not single-valued: it is a positively homogeneous set-valued map (i.e., one whose graph is a cone) which can also be transposed. We shall then prove two things: first, that for any convex cone $Q \subset W(t, z)$,

$$(r(T-t, t)Q)^+ = r(T-t, t)^{*-1}(Q^+), \tag{1.10}$$

and second, that the transpose $r(T-t, t)^*$ can be computed in the following way:

$$r(T-t, t)^* \pi = q(t) \tag{1.11}$$

where q is a solution to the system (1.6), (1.8) satisfying $q(T) = \pi$. By piecing together all this information, we obtain the existence of a solution q of (1.6)-(1.9) satisfying

$$q(T) \in T_{R(T,C)}(z(T))^+, \tag{1.12}$$

$$q(t) \in W(t, z)^+ \quad \text{for all } t \in [0, T[. \tag{1.13}$$

It also implies the following invariance property of reachable sets:

(1.14) If $T_{R(T,C)}(z(T))^+ \neq \{0\}$, then for all $t \in [0, T]$, $T_{R(t,C)}(z(t))^+ \neq \{0\}$.

This result is of the same nature as a theorem of Ważewski stating that the boundary point of reachable set can be reached only by a boundary trajectory.

The inclusions (1.12)-(1.13) are additional information described via *reachable sets*. For nonlinear systems the reachable sets and, consequently, the set of variations $W(t, z)$ are not known a priori. But condition (1.13) still allows us to eliminate some candidates for optimality among those satisfying the maximum principle. Let us emphasize that it is enough to know *one* element $w \in W(t, z)$, such that the solution q of (1.6), (1.7) satisfies $\langle q(t), w\rangle < 0$, to deduce that z is not optimal.

Inclusion (1.13) can also be seen as a *higher-order* optimality condition, since it deals with variations of $z(\cdot)$ of all orders. High-order necessary conditions involving higher-order derivatives of g are (of course) of an entirely different nature.

The high-order necessary conditions in optimization have two features:

(1) Necessary conditions involving the high-order variations of constraints;

(2) Calculus of high-order variations.

Here we shall not divide any calculus of sets $W(t, z)$. In [19] the interested reader can find many examples of variations corresponding to piecewise C^∞-controls. They are constructed via Lie brackets of some vector fields. However, because of the Lavrentiev phenomenon, we should not expect such regularity of optimal trajectories. Still the results of [19] can be used at *regular* enough points of optimal control. The irregular points are much more difficult to address and require further investigations.

We shall study a more general dynamical system than the parametrized control system (1.2), (1.3), the so-called differential inclusion

$$x' \in F(x). \tag{1.15}$$

This is a generalized differential equation to which the control system (1.2), (1.3) can be reduced by setting $F(x) = f(x, U)$. When f is continuous, the Filippov Theorem (see [1, p. 91]) says that the solutions of (1.15) and (1.2), (1.3) do coincide.

In general the set-valued map F cannot be parametrized in a way that reduces the system (1.15) to (1.2), (1.3). The main reason for this is the restriction on admissible controls (1.3). Still this can be done when F has *convex* compact images and is continuous in the Hausdorff metric. But even in this case the parametrization would only be continuous, and therefore would not be very useful because of the lack of differentiability of f.

The differential inclusions, besides being a description of more general dynamical systems, provide a mathematical tool for studying nonsmooth control systems, closed-loop control systems:

$$x' = f(x, u), \qquad u \in U(x), \tag{1.16}$$

and implicit dynamical systems

$$f(x, x') = 0. \tag{1.17}$$

We refer to [1], [9], [22], [6], and references therein for the corresponding examples of systems whose models are described by (1.16), (1.17).

Setting $F(x) = \bigcup_{u \in U(x)} f(x, u)$ and $F(x) = \{v: f(x, v) = 0\}$ we reduce (1.16) and (1.17), respectively, to the differential inclusion (1.15).

Recall that the dynamical system (1.17) appears in the Lagrange problem (see [28]). Two ways to treat (1.17) are described in [28]. One is an unjustified multiplier rule. The second is (again) an unjustified assumption that (1.17) can be rewritten as a control system (1.2), (1.3). In this paper we treat (1.17) via differential inclusion techniques.

Properties of the dynamical system given by (1.15) depend on the graph of the set-valued map F.

Actually the generalized differential equation (1.15) inherits many features of ODEs (see [1]). The one we exploit the most here is the variational inclusion, which is as useful as variational equations arising in ODEs. It has been extended to variational inclusions in [13], [12], and independently in [23]. Many results concerning inclusions can be found in [1], [9]–[16], [18], [23] (see also the references therein).

The maximum principle for differential inclusions has been proved in [9], [10], [12], [18], [23]. It involves graphical derivatives of the set-valued map F ([12], [23]), generalized Jacobians of selections for F [18], or generalized gradients of Hamiltonians [9], [10]:

$$H(x, p) = \sup \{\langle p, e\rangle : e \in F(x)\}.$$

We prefer the "graphical" approach mainly for two reasons:

(1) In general, even for smooth control systems, H is merely Lipschitz. Hence we are led to differentiate H in one generalized way or another. There is not yet any convenient notion of higher-order generalized derivatives of H adequate for our purposes. Neither is it clear how we can solve the nonsmooth Hamiltonian inclusions. Rather, we deal with convex subcones of tangent cones to graph (F) and the associated convex processes. The convex process is a set-valued analogue of linear operators (see [25], [2]). In particular, the Kalman rank condition can be extended to convex processes [3].

(2) In the examples of applications we provide here, the Hamiltonian maximum principle is less powerful than that involving the adjoint system (see § 4, Remark 4.10 for a detailed discussion).

Tangent vectors to reachable sets are studied via local variations in § 2. In § 3 we investigate the adjoint of the reachable process, $r(T-t, t)^*$. The cone $T_{R(T,C)}(z(T))^+$ is studied in § 4. Section 5 is devoted to necessary conditions for problem (1.1) for the (usual) control system (1.2), (1.3), the closed-loop control system (1.16), and the implicit dynamical system (1.17). In § 6 we sketch how the same approach can be used to study control systems with jumps (deterministic impulse control systems). Examples are provided in § 7.

We do not present here a thorough study of high-order variation. Many results concerning smooth cases can be found in [19]. In the more general framework (1.15) we deal with the extended notion of Lie brackets for set-valued maps. A second-order result can be found in [14]. However, the higher-order variations require a further investigation.

2. Tangent vectors to reachable sets. One of the main tools we use here is the following result due to Filippov [11].

THEOREM (Filippov). *Let $y:[a, b]\to\mathbb{R}^n$ be an absolutely continuous function and $G:[a, b]\times\mathbb{R}^n\to\mathbb{R}^n$ be a set-valued map with closed images such that:*

(i) *For all $x\in\mathbb{R}^n$, the map $t\to G(t, x)$ is measurable;*

(ii) *For some $\varepsilon>0$, $k\in L^1(a, b)$ and all t, $G(t, \cdot)$ has nonempty images and is $k(t)$-Lipschitz on $y(t)+\varepsilon B$.*

Set $K=\exp(\int_a^b k(t)\,dt)$, $\rho:=\int_a^b \operatorname{dist}(y'(t), G(t, y(t)))\,dt$. If $\rho<\varepsilon/K$, then there exists an absolutely continuous function $x:[a, b]\to\mathbb{R}^n$ satisfying $x(a)=y(a)$,

$$x'(s)\in G(s, x(s)) \quad \text{a.e. in } [a, b],$$

$\|x-y\|_{C(a,b)}\leqq K\rho$ and for almost all $t\in[a, b]$

$$\|x'(t)-y'(t)\|\leqq k(t)\rho\exp\left(\int_a^t k(s)\,ds\right)+\operatorname{dist}(y'(t), G(t, y(t))).$$

Remark. The proof can also be found in [1] under an additional assumption that G is continuous in t. In [9, p. 115] the theorem above is stated in a weaker form, but the proof allows us to deduce the above stronger version. We provide a sketch of such a deduction. The function x is constructed as the limit of a Cauchy sequence $x_i \in C(a, b; \mathbb{R}^n)$ $i = 0, 1, \cdots$ of absolutely continuous functions satisfying $x_i(a) = y(a)$ and for almost all $t \in [a, b]$ and all $i \geqq 1$:

$$\|x_{i+1}'(t) - x_i'(t)\| \leqq k(t)\|x_i(t) - x_{i-1}(t)\| \leqq k(t)\rho \frac{(\int_a^t k(s)\, ds)^{i-2}}{(i-2)!},$$

$$\|x_1'(t) - y'(t)\| = \operatorname{dist}(y'(t), G(t, y(t))).$$

Hence for almost all $t \in [a, b]$ the sequence $\{x_i'(t)\}$ is also Cauchy. This and Lebesgue's dominated convergence theorem yield the existence of $x \in C(a, b)$ such that for all $t \in [a, b]$

$$x(t) = x(a) + \int_a^t \lim_{i \to \infty} x_i'(s)\, ds.$$

Hence x is absolutely continuous and we finally obtain that

$$x_i'(s) \to x'(s) \quad \text{a.e. in } [a, b].$$

Moreover, for almost all $t \in [a, b]$

$$\begin{aligned}\|x_{i+1}'(t) - y'(t)\| &\leqq \sum_{j=1}^{i} \|x_{j+1}'(t) - x_j'(t)\| + \|x_1'(t) - y'(t)\| \\ &\leqq k(t)\rho \sum_{j=0}^{i} \left(\int_a^t k(s)\, ds\right)^j \Big/ j! + \|x_1'(t) - y'(t)\| \\ &\leqq k(t)\rho \exp\left(\int_a^t k(s)\, ds\right) + \operatorname{dist}(y'(t), G(t, y(t))).\end{aligned}$$

Taking the limit we obtain that for almost all $t \in [a, b]$

$$\|x'(t) - y'(t)\| \leqq k(t)\rho \exp\left(\int_a^t k(s)\, ds\right) + \operatorname{dist}(y'(t), G(t, y(t))).$$

Consider a set-valued map F from $\mathbb{R}^n$ to $\mathbb{R}^n$ and a differential inclusion

$$x' \in F(x). \tag{2.1}$$

A function $x \in W^{1,1}(0, T)$, $T > 0$ (the Sobolev space) is called a trajectory of (2.1) if for almost all $t \in [0, T]$, $x'(t) \in F(x(t))$. We denote by S_t the set of all trajectories of (2.1) defined on the time interval $[0, t]$. The reachable set of the inclusion (2.1) from a point $\xi \in \mathbb{R}^n$ at time $t \geqq 0$ is given by

$$R(t, \xi) = \{x(t) : x \in S_t, x(0) = \xi\}.$$

We observe that the reachable sets enjoy the semigroup property:

$$\begin{aligned} &R(t + h, \xi) = R(t, R(h, \xi)) \quad \text{for all } t, h \geqq 0, \\ &R(0, \xi) = \xi. \end{aligned} \tag{2.2}$$

Let $z \in S_T$ be a given trajectory. In this section we study tangent vectors to the reachable set $R(T, C)$ at $z(T)$. We call a set $Q \subset \mathbb{R}^n$ a cone if for all $\lambda \geqq 0$, $\lambda Q \subset Q$. First recall the following definition.

DEFINITION 2.1. Let K be a subset of $\mathbb{R}^n$ and $x \in K$. The (Bouligand) contingent cone to K at x is given by

$$T_K(x) = \{v \in \mathbb{R}^n : \exists h_i \to 0+,\ v_i \to v \text{ such that } x + h_i v_i \in K\}.$$

The intermediate tangent cone to K at x is defined by

$$I_K(x) = \{v \in \mathbb{R}^n : \forall h_i \to 0+ \exists v_i \to v \text{ such that } x + h_i v_i \in K\}.$$

We refer to [2], [12] for properties of $T_K(x)$, $I_K(x)$. Throughout the entire paper we assume that the set-valued map F in the right-hand side of the differential inclusion (2.1) satisfies the following assumption:

$$(H_1) \quad \begin{cases} \text{Dom } F := \{x: F(x) \neq \emptyset\} \text{ is open,} \\ F \text{ has compact images and is Lipschitzian on Dom } F. \end{cases}$$

DEFINITION 2.2. Let $F: \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map that is locally Lipschitzian at x and $y \in F(x)$. The derivative of F at (x, y) is the set-valued map $dF(x, y): \mathbb{R}^n \rightrightarrows R^n$ given by the following: for all $u \in \mathbb{R}^n$,

$$v \in dF(x, y)u \Leftrightarrow \lim_{h \to 0+} \operatorname{dist}\left(v, \frac{F(x+hu)-y}{h}\right) = 0.$$

Observe that graph $(dF(x, y)) := \{(u, v): v \in dF(x, y)u\}$ is a closed cone equal to the intermediate tangent cone to graph (F) at (x, y). We refer to [12]-[14] for some properties and applications of the set-valued derivative.

We denote by co F the convexified set-valued map, i.e., for all $x \in \mathbb{R}^n$, co $F(x)$ is the convex hull of $F(x)$.

Consider the "linearized inclusion"

$$w'(s) \in d \text{ co } F(z(s), z'(s))w(s) \quad \text{a.e.} \tag{2.3}$$

For all h, $t \geqq 0$, $\xi \in \mathbb{R}^n$ define the reachable set $r(h, t)\xi$ of (2.3) by

$$r(h, t)\xi = \{w(t+h): w \in W^{1,1}(t, t+h) \text{ satisfies (2.3)}, w(t) = \xi\}.$$

DEFINITION 2.3. Let $t \in [0, T[$. Set

$$W(t, z) = \{v : \exists h_i \geqq 0,\ \mu_i \to 0+ \text{ such that } \lim_{i \to \infty} h_i = 0,\ z(t+h_i) + \mu_i v \in R(t+h_i, C) + o(\mu_i)B\},$$

$$\mathcal{W}(t, z) = \{v : \forall \mu_i \to 0+ \exists h_i \to 0,\ h_i \geqq 0 \text{ such that } z(t+h_i) + \mu_i v \in R(t+h_i, C) + o(\mu_i)B\}.$$

Observe that $W(t, z)$ and $\mathcal{W}(t, z)$ are closed cones. Moreover, for all $t \in [0, T[$

$$T_{R(t,C)}(z(t)) \subset W(t, z), \qquad I_{R(t,C)}(z(t)) \subset \mathcal{W}(t, z) \subset W(t, z) \tag{2.4}$$

and, in particular, $T_C(z(0)) \subset W(0, z)$.

Remark. When for some integer $k \geqq 1$, $\mu_i = h_i^k$, the vector v can be seen as the kth-order variation of $R(\cdot)$ at (t, z).

Actually, variations of $R(\cdot, C)$ at (t, z) are mapped by $r(T-t, t)$ into the tangent vectors to $R(T, C)$.

THEOREM 2.4. *Assume that* (H_1) *is verified and let* $t \in [0, T[$. *Then for all* $t < \tau < T$

$$r(\tau - t, t)W(t, z) \subset T_{R(\tau,C)}(z(\tau)),\ r(\tau - t, t)\mathcal{W}(t, z) \subset I_{R(\tau,C)}(z(\tau)),$$

$$r(T - t, t)T_{R(t,C)}(z(t)) \subset T_{R(T,C)}(z(T)).$$

To prove the theorem above, we need a consequence of the Filippov-Ważewski relaxation result (see [1, p. 124]):

Consider the convexified inclusion

$$\begin{cases} x'(s) \in \operatorname{co} F(x(s)) & \text{a.e.,} \\ x(0) \in C. \end{cases} \tag{2.5}$$

PROPOSITION 2.5. *Assume that* (H_1) *holds true. Then for all* $t \in [0, T]$ *the contingent (respectively, intermediate) cones to the reachable sets of* (2.1) *and* (2.5) *at time* t *taken at the point* $z(t)$ *do coincide.*

Proof of Theorem 2.4. By Proposition 2.5, we may asume that F has convex images. Fix a solution w of (2.3) and let $h_i \geqq 0$, $\mu_i \to 0+$, $v_i \to v = w(t)$ be such that $\lim_{i\to\infty} h_i = 0$, $z(t+h_i) + \mu_i v_i \in R(t+h_i, C)$. For all $s \in [t+h_i, \tau]$ set

$$y_i(s) = z(s) + \mu_i \left(v_i + \int_{t+h_i}^{s} w'(p)\, dp \right)$$

and let $L \geqq 1$ denote the Lipschitz constant of F. Then for almost all $s \in [t+h_i, \tau]$ and all large i

$$\begin{aligned} \operatorname{dist}(y_i'(s), F(y_i(s))) &\leqq \operatorname{dist}(z'(s) + \mu_i w'(s), F(z(s) + \mu_i w(s))) \\ &\quad + L\mu_i \left(\|v_i - v\| + \int_t^{t+h_i} \|w'(p)\|\, dp \right) \\ &\leqq L\mu_i \left(\|w'(s)\| + \|w(s)\| + \|v_i - v\| + \int_t^{t+h_i} \|w'(p)\|\, dp \right). \end{aligned} \tag{2.6}$$

Moreover,

$$\lim_{i\to\infty} \left(\|v_i - v\| + \int_t^{t+h_i} \|w'(p)\|\, dp \right) = 0$$

and, by definition of dF, for almost all $s \in [t, \tau]$

$$\lim_{i\to\infty} \operatorname{dist}(z'(s) + \mu_i w'(s), F(z(s) + \mu_i w(s)))/\mu_i = 0.$$

Thus, by the Lebesgue dominated convergence theorem and (2.6),

$$\lim_{i\to\infty} \int_{t+h_i}^{\tau} \operatorname{dist}(y_i'(s), F(y_i(s)))\, ds/\mu_i = 0.$$

By the Filippov Theorem there exist

$$r_i \in R(\tau - t - h_i, z(t+h_i) + \mu_i v_i) \subset R(\tau, C)$$

such that $\|r_i - y_i(\tau)\| = o(\mu_i)$.

Since

$$\lim_{i\to\infty} (y_i(\tau) - z(\tau))/\mu_i = \lim_{i\to\infty} \left(v_i + \int_{t+h_i}^{\tau} w'(p)\, dp \right) = w(\tau),$$

we end the proof. □

THEOREM 2.6. *Assume that* (H_1) *is verified and let* $0 \leqq t \leqq \tau \leqq T$. *Then the set*

$$\{(w(t), w(\tau)) : w(t) \in T_{R(t,C)}(z(t)),\ w \in W^{1,1}(t, \tau) \text{ is a trajectory of } (2.3)\}$$

is contained in

$$T_{\{(x,y):x\in R(t,C),y\in R(\tau-t,x)\}}(z(t), z(\tau)).$$

Proof. By the proof of Theorem 2.4 in the case when $h_i = 0$ for all $i \geqq 1$, we know that there exist $\mu_i \to 0+$, $v_i \to v$, $r_i \in R(\tau - t, z(t) + \mu_i v_i)$ such that $z(t) + \mu_i v_i \in R(t, C)$ and $\|r_i - z(\tau) - \mu_i(v_i + \int_t^\tau w'(p)\, dp)\| = o(\mu_i)$. Hence

$$\lim_{i\to\infty} (z(t) + \mu_i v_i - z(t), r_i - z(\tau))/\mu_i = \left(v, v + \int_t^\tau w'(p)\, dp\right) = (w(t), w(\tau)).$$

It is shown in [16] that under the hypothesis (H_1) the reachable map R has the following (first-order) expansion: for all ξ near $z(t)$ and all small $h > 0$

$$R(h, \xi) = \xi + h \text{ co } F(z(t)) + o(t, h) \tag{2.7}$$

where

$$\lim_{h\to 0+, \xi \to z(t)} \|o(t, h)\|/h = 0$$

and the equality in (2.7) must be understood in the following way:

$$R(h, \xi) \subset \xi + h \text{ co } F(z(t)) + o(t, h)B,$$

$$\xi + h \text{ co } F(z(t)) \subset R(h, \xi) + o(t, h)B.$$

On the other hand, the function $z(\cdot)$ being absolutely continuous, for almost all $t \in [0, T]$ and all $h > 0$ we can write $z(t + h) = z(t) + hz'(t) + o(h)$. Applying (2.7) with $\xi = z(t)$ and using Definition 2.3 we obtain

$$\text{co } F(z(t)) - z'(t) \subset \mathcal{W}(t, z) \quad \text{a.e. in } [0, T]. \tag{2.8}$$

We have even a stronger result that we shall use in Theorem 2.9.

THEOREM 2.7. *Assume that* (H_1) *holds true. Then* $\mathcal{W}(t, z) + T_{R(t,C)}(z(t)) \subset W(t, z)$, $\mathcal{W}(t, z) + I_{R(t,C)}(z(t)) = \mathcal{W}(t, z)$.

Proof. Fix $w \in \mathcal{W}(t, z)$, $v \in T_{R(t,C)}(z(t))$ and let $\mu_i \to 0+$, $v_i \to v$ be such that $z(t) + \mu_i v_i \in R(t, C)$. Fix $h_i \to 0+$, $w_i \to w$, $y_i \in S_{t+h_i}$ such that $z(t + h_i) + \mu_i w_i \in R(h_i, z(t))$, $y_i(t) = z(t)$, $y_i(t + h_i) = z(t + h_i) + \mu_i w_i$. Set $\bar{y}_i = y_i + \mu_i v_i$. Then dist $(\bar{y}_i'(s), F(\bar{y}_i(s))) \leqq$ dist $(y_i'(s), F(y_i(s))) + L\mu_i\|v_i\| = L\mu_i\|v_i\|$, where L denotes the Lipschitz constant of F. This and Filippov's Theorem imply the existence of $x_i \in S_{t+h_i}$ such that $x_i(t) = \bar{y}_i(t) = z(t) + \mu_i v_i \in R(t, C)$,

$$\begin{aligned} x_i(t + h_i) &= \bar{y}_i(t + h_i) + o(\mu_i) \\ &= z(t + h_i) + \mu_i(w_i + v_i) + o(\mu_i) \in R(h_i, x_i(t)) \subset R(h_i, R(t, C)). \end{aligned}$$

Hence, from (2.2),

$$z(t + h_i) + \mu_i(w_i + v_i) \in R(t + h_i, C) + o(\mu_i).$$

Definition 2.3 ends the proof of the first statement. The proof of the second one is analogous; therefore we omit it. □

In § 4 we study "normal" cones to reachable sets along the trajectory z via a duality technique applied to convex subcones of the set $W(t, z)$. Next we introduce an example of such a subcone.

DEFINITION 2.8. Let $t \in [0, T]$. A vector $v \in \mathbb{R}^n$ is called a smooth variation of order $k > 0$ at (t, z) if

$$\lim_{\substack{h \to 0+ \\ t' \to t+}} \operatorname{dist}\left(v, \frac{R(h, z(t')) - z(t'+h)}{h^k}\right) = 0.$$

The set of all variations of order k is denoted by $R^k(t, z)$. The closed cone spanned by all variations is called the expansion cone of the reachable map at (t, z) and is denoted by $R^\infty(t, z)$:

$$R^\infty(t, z) = \operatorname{cl} \bigcup_{\substack{\lambda \geqq 0 \\ k > 0}} \lambda R^k(t, z).$$

The expansion cone at a stationary trajectory is introduced in [14] for studying the problem of local controllability at a point of equilibrium. Clearly, whenever $v \in R^k(t, z)$, for all $\mu_i \to 0+$ there exist $h_i \to 0+$ such that $z(t+h_i) + \mu_i v \in R(h_i, z(t)) + o(\mu_i)$. Hence Theorem 2.7 yields $T_{R(t,C)}(z(t)) + R^k(t, z) \subset W(t, z)$. Moreover,

$$\begin{aligned} &I_{R(t,C)}(z(t)) + R^\infty(t, z) \subset \mathcal{W}(t, z), \\ &T_{R(t,C)}(z(t)) + R^\infty(t, z) \subset W(t, z). \end{aligned} \tag{2.9}$$

THEOREM 2.9. *Assume that* (H_1) *holds true. Then* $R^\infty(t, z)$ *is a closed convex subcone of the cone of variations* $\mathcal{W}(t, z)$ *satisfying* (2.9).

This result is an immediate consequence of the closedness of $\mathcal{W}(t, z)$ and Lemma 2.10.

LEMMA 2.10. *If* (H_1) *holds true then we have the following*:

(i) *For all* $K > k$, $0 \in R^k(t, z) \subset R^K(t, z)$;

(ii) *For all* $k > 0$, $(n+1)^{-k} \operatorname{co} R^k(t, z) \subset R^k(t, z)$.

Proof. Clearly, for all $k > 0$

$$0 \in R^k(t, z). \tag{2.10}$$

Fix $K > k > 0$ and observe that for all $v \in \mathbb{R}^n$, $t' \in [0, T[$, $h \in]0, 1[$ we have $h^{K/k} < h$ and

$$\begin{aligned} &\operatorname{dist}\left(v, \frac{R(h, z(t')) - z(t'+h)}{h^K}\right) \\ &\quad \leqq \operatorname{dist}\left(v, \frac{R(h^{K/k}, z(t'+h-h^{K/k})) - z(t'+h-h^{K/k}+h^{K/k})}{(h^{K/k})^k}\right). \end{aligned}$$

This and Definition 2.8 imply (i). To prove (ii) fix $k > 0$, $\lambda_i \geqq 0$, $v_i \in R^k(t, z)$, $i = 0, \cdots, m$ satisfying $\sum_{i=0}^m \lambda_i = 1$. We claim that

$$\sum_{i=0}^m \lambda_i^k v_i \in R^k(t, z). \tag{2.11}$$

Indeed, consider $t_j \to t+$, $h_j \to 0+$. Then

$$z(t_j + \lambda_0 h_j) + h_j^k \lambda_0^k v_0 \in R(\lambda_0 h_j, z(t_j)) + o(h_j^k) B$$

where $\lim_{j \to \infty} o(h_j^k)/h_j^k = 0$. We proceed by the induction. Assume that we have already proved that for some $0 \leqq s < n$ and all j

$$z\left(t_j + h_j \sum_{i=0}^s \lambda_i\right) + h_j^k \sum_{i=0}^s \lambda_i^k v_i \in R\left(h_j \sum_{i=0}^s \lambda_i, z(t_j)\right) + o(h_j^k) B \tag{2.12}$$

with $\lim_{j\to\infty} o(h_j^k)/h_j^k = 0$. By Definition 2.8 applied with $t' = t_j + h_j \sum_{i=0}^{s} \lambda_i$, $h = \lambda_{s+1} h_j$

$$z(t' + \lambda_{s+1} h_j) + h_j^k \lambda_{s+1}^k v_{s+1} \in R(h_j \lambda_{s+1}, z(t')) + o(h_j^k) B.$$

This and the Filippov Theorem yield

$$z(t' + \lambda_{s+1} h_j) + h_j^k \sum_{i=0}^{s+1} \lambda_i^k v_i \in R\left(h_j \lambda_{s+1}, z(t') + h_j^k \sum_{i=0}^{s} \lambda_i^k v_i\right) + o(h_j^k) B \subset \text{(by (2.12))},$$

$$R\left(h_j \lambda_{s+1}, R\left(h_j \sum_{i=0}^{s} \lambda_i, z(t_j)\right)\right) + o(h_j^k) B = R\left(h_j \sum_{i=0}^{s+1} \lambda_i, z(t_j)\right) + o(h_j^k) B.$$

Hence (2.12) is valid also with s replaced by $s+1$. Applying (2.12) with $s = m$, we obtain that

$$\lim_{j\to\infty} \operatorname{dist}\left(\sum_{i=0}^{m} \lambda_i^k v_i, \frac{R(h_j, z(t_j)) - z(t_j + h_j)}{h_j^k}\right) = 0,$$

and since $\{t_j\}$ and $\{h_j\}$ are arbitrary, Definition 2.8 implies (2.11). On the other hand, by the Carathéodory Theorem for all $v \in \operatorname{co} R^k(t, z)$ there exist $\mu_i \geqq 0$, $v_i \in R^k(t, z)$ such that $\sum_{i=0}^{n} \mu_i = 1$ and $\sum_{i=0}^{n} \mu_i v_i = v$. Observe that $\sum_{i=0}^{n} \sqrt[k]{\mu_i}/(n+1) \leqq 1$. Applying (2.11) with

$$\lambda_i = \sqrt[k]{\mu_i}/(n+1), \quad v_{n+1} = 0, \quad \lambda_{n+1} = 1 - \sum_{i=0}^{n} \sqrt[k]{\mu_i}/(n+1),$$

we obtain that $(n+1)^{-k} v = \sum_{i=0}^{n+1} \lambda_i^k v_i \in R^k(t, z)$. This proves (ii). □

3. The adjoint process $r(T-t, t)^*$. Recall that for a subset K of a Banach space E, its positive polar cone is given by

$$K^+ = \{p \in E^* : \forall u \in K, \langle p, u\rangle \geqq 0\}.$$

We also recall the following definition.

DEFINITION 3.1. A set-valued map $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is called a (closed) convex process if graph (G) is a closed convex cone.

We refer to Rockafellar [25], who introduced and studied this notion, and to Aubin and Ekeland [2] for further properties.

DEFINITION 3.2. Let $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ be a set-valued map. The adjoint map $G^* : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is given by $p \in G^*(q)$ if and only if for all $(x, y) \in \text{graph } (G)$, $\langle p, x\rangle \leqq \langle q, y\rangle$. In other words $p \in G^*(q) \Leftrightarrow (-p, q) \in \text{graph } (G)^+$.

Observe that the adjoint G^* is a closed convex process.

Let $\{A(s) : s \in [0, T]\}$ be a given family of closed convex processes from $\mathbb{R}^n$ to $\mathbb{R}^n$ satisfying the following:

(H_2) (i) For all $w \in \mathbb{R}^n$ the map $s \to A(s)w$ is measurable.
(ii) For all $s \in [0, T]$, the map $w \to A(s)w$ is $k(s)$-Lipschitzian, where $k \in L^\infty(0, T)$.

For all $0 \leqq t \leqq \tau \leqq T$, we investigate the adjoint $r(\tau - t, t)^*$ by studying the inclusions

$$w'(s) \in A(s)w(s) \quad \text{a.e.}, \tag{3.1}$$

$$-q'(s) \in A(s)^* q(s) \quad \text{a.e.} \tag{3.2}$$

in the case when

(H_3) $\qquad \text{graph } (A(s)) \subset \text{graph } (d \text{ co } F(z(s), z'(s)))$ a.e. in $[0, T]$.

For a subset $Q \subset \mathbb{R}^n$ we denote by $r_Q(\tau - t, t)$ the restriction of r to Q, i.e.,

$$r_Q(\tau - t, t)x = \begin{cases} r(\tau - t, t)x & \text{when } x \in Q, \\ \varnothing & \text{otherwise.} \end{cases}$$

The main result of this section is Theorem 3.3

THEOREM 3.3. *If a family* $\{A(s)\colon s \in [0, T]\}$ *of closed convex processes from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$ *satisfies* (H_2) *and* (H_3), *then for all* $b \in \mathbb{R}^n$, *convex cone* $Q \subset \mathbb{R}^n$, *and* $0 \leqq t \leqq \tau \leqq T$:

(a) $r(\tau - t, t)^* b \subset \{q(t)\colon q \in W^{1,\infty}(t, \tau)$ *satisfies* (3.2), $q(\tau) = b\}$;

(b) $r_Q(\tau - t, t)^* b \subset \{q(t)\colon q \in W^{1,\infty}(t, \tau)$ *satisfies* (3.2), $q(\tau) = b\} - Q^+$;

(c) $(r(\tau - t, t)Q)^+ \subset \{q(\tau)\colon q \in W^{1,\infty}(t, \tau)$ *satisfies* (3.2), $q(t) \in Q^+\}$.

To prove the above theorem we associate with all $0 \leqq t \leqq \tau \leqq T$ the convex process $\hat{r}(\tau - t, t)\colon \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined by the following: for all $\xi \in \mathbb{R}^n$,

$$\hat{r}(\tau - t, t)\xi = \{w(\tau)\colon w \text{ satisfies (3.1) on } [t, \tau],\ w(t) = \xi\}. \tag{3.3}$$

Then by the definition of the adjoint map, for all $b \in R^n$

$$r(\tau - t, t)^* b \subset \hat{r}(\tau - t, t)^* b, \tag{3.4}$$

$$r_Q(\tau - t, t)^* b \subset \hat{r}_Q(\tau - t, t)^* b, \tag{3.5}$$

$$(r(\tau - t, t)Q)^+ \subset (\hat{r}(\tau - t, t)Q)^+. \tag{3.6}$$

Theorem 3.3 follows from the inclusions above and the following two lemmas.

LEMMA 3.4. *If* (H_2) *holds true, then for any* $0 \leqq t \leqq \tau \leqq T$ *and* $b \in \mathbb{R}^n$

$$\hat{r}(\tau - t, t)^* b = \{q(t)\colon q \in W^{1,\infty}(t, \tau) \text{ satisfies (3.2)}, q(\tau) = b\}. \tag{3.7}$$

LEMMA 3.5. *If* (H_2) *holds true, then for any convex cone* $Q \subset \mathbb{R}^n$ *and* $b \in$ Dom $\hat{r}(\tau - t, t)^*$ *we have the following*:

$$\hat{r}_Q(\tau - t, t)^* b = \hat{r}(\tau - t, t)^* b - Q^+, \tag{3.8}$$

$$(\hat{r}(\tau - t, t)Q)^+ = \hat{r}(\tau - t, t)^{*-1}(Q^+).$$

Proof of Lemma 3.4. Fix $0 \leqq t \leqq \tau \leqq T$. Let us set:

$X = W^{1,2}(t, \tau)$, $Y = L^2(t, \tau) \times L^2(t, \tau)$;

$L = \{(x, y) \in Y\colon y(s) \in A(s)x(s)$ a.e. in $[t, \tau]\}$;

D, the differential operator on X, $Dx = x'$;

γ, the trace operator on X, $\gamma(x) = (x(t), x(\tau))$.

Observe that L is a closed convex cone and, by the measurable selection theorem (see [26]),

$$L^+ = \{(-p, q) \in Y^*\colon p(s) \in A(s)^* q(s) \text{ a.e. in } [t, \tau]\}. \tag{3.9}$$

We claim that

$$\operatorname{Im}(1 \times D) - L = Y. \tag{3.10}$$

To prove it we must verify that for all $(u, v) \in Y$ there exists $x \in X$ satisfying

$$x'(s) \in A(s)(x(s) - u(s)) + v(s) \quad \text{a.e. in } [t, \tau]. \tag{3.11}$$

Fix $(u, v) \in Y$ and observe that, by (H_2), the set-valued map $[t, \tau] \times \mathbb{R}^n \ni (s, x) \to A(s)(x - u(s)) + v(s)$ is measurable in s, and for almost all s it is Lipschitzian in x with the Lipschitz constant $k(s)$. Moreover, $\operatorname{dist}(0, A(s)(-u(s)) + v(s)) \leqq k(s)\ \|u(s)\| + \|v(s)\|$. By the Filippov Theorem there exist $M \geqq 0$ and $x \in W^{1,1}(t, \tau)$ satisfying (3.11) and such that

$$\|x'(s)\| \leqq Mk(s) + k(s)\|u(s)\| + \|v(s)\| \quad \text{a.e. in } [t, \tau].$$

Thus $\|x'\| \in L^2(t, \tau)$ and therefore $x \in X$. Hence we have proved (3.10). By Lemma 1.3 of [3] and (3.10) we obtain

$$((1 \times D)^{-1} L)^+ = (1 \times D)^*(L^+). \tag{3.12}$$

Clearly, $\gamma((1 \times D)^{-1} L) \subset \text{graph}\,(\hat{r}(\tau - t, t))$ and by (3.12), $\gamma^* \,\text{graph}\,(\hat{r}(\tau - t, t))^+ \subset ((1 \times D)^{-1} L)^+ = (1 \times D)^*(L^+)$. Hence for all $(a, b) \in \text{graph}\,(\hat{r}(\tau - t, t))^+$ there exists $(-p, q) \in L^+$ such that

$$\gamma^*(a, b) = (1 \times D)^*(-p, q). \tag{3.13}$$

This implies that for all $w \in W_0^{1,2}(t, \tau)$,

$$0 = \langle (1 \times D)^*(-p, q), w \rangle = \int_t^\tau (-pw + qw')(s)\, ds = \int_t^\tau w'(s) \left(q(s) + \int_t^s p(r)\, dr \right) ds.$$

Thus, $q \in W^{1,2}(t, \tau)$ and $q' = -p$. By (3.9), $-q'(s) \in A(s)^* q(s)$ almost everywhere in $[t, \tau]$. From Proposition 1.7b of [3] we deduce that $q \in W^{1,\infty}(t, \tau)$. Moreover by (3.13) for all $x \in X$, $\langle (a, b), (x(t), x(\tau)) \rangle = \langle (q', q), (x, x') \rangle = q(\tau)x(\tau) - q(t)x(t)$. Hence $(-a, b) = (q(t), q(\tau))$, and $q(t) \in \hat{r}(\tau - t, t)^* q(\tau)$. We have proved that for all $b \in \mathbb{R}^n$, $\hat{r}(\tau - t, t)^* b$ is contained in the right-hand side of (3.7). On the other hand, if q satisfies (3.2) then for all solutions w of (3.1)

$$q(\tau)w(\tau) - q(t)w(t) = \langle (q', q), (w, w') \rangle \geqq 0.$$

This yields that $q(t) \in r(\tau - t, t)^* q(\tau)$ and ends the proof. □

To prove Lemma 3.5 we apply some results from [2, pp. 142-143] concerning closed convex processes. Since in general $\hat{r}(\tau - t, t)$ is not closed we need the following lemma.

LEMMA 3.6. *If* (H_2) *holds true then* $\hat{r}(\tau - t, t)$ *is Lipschitzian on* $\mathbb{R}^n$ *and the set-valued map* cl $\hat{r}(\tau - t, t)$ *defined as follows: for all* $u \in \mathbb{R}^n$, cl $\hat{r}(\tau - t, t)u = \overline{\hat{r}(\tau - t, t)u}$ *is a Lipschitzian on* $\mathbb{R}^n$ *closed convex process. Moreover,* $(\text{cl}\,\hat{r}(\tau - t, t))^* = \hat{r}(\tau - t, t)^*$ *is an upper semicontinuous set-valued map with compact images mapping bounded sets to bounded sets and Dom* $\hat{r}(\tau - t, t)^* = \hat{r}(\tau - t, t)(0)^+$.

Proof of Lemma 3.6. Since $0 \in \hat{r}(\tau - t, t)0$, the set $\hat{r}(\tau - t, t)0$ is nonempty. Fix any $u \in \mathbb{R}^n$ such that $\hat{r}(\tau - t, t)u \neq \varnothing$ and let w be a solution of (3.1) on $[t, \tau]$ satisfying $w(t) = u$. Pick $v \in \mathbb{R}^n$ and set $y(\cdot) = w(\cdot) + v - u$. Then $\text{dist}\,(y'(s), A(s)y(s)) = \text{dist}\,(w'(s), A(s)(w(s) + v - u)) \leqq k(s)\|v - u\|$. This and the Filippov Theorem imply the existence of a solution $\bar{w}$ of (3.1) defined on $[t, \tau]$ and satisfying $\bar{w}(t) = y(t) = w(t) + v - u = v$,

$$\|\bar{w}(\tau) - y(\tau)\| \leqq M\|v - u\|$$

where M does not depend on v, u. Thus $\hat{r}(\tau - t, t)v \neq \varnothing$ and

$$\|\bar{w}(\tau) - w(\tau)\| \leqq \|\bar{w}(\tau) - y(\tau)\| + \|y(\tau) - w(\tau)\| \leqq M\|v - u\| + \|v - u\|,$$

i.e., $\hat{r}(\tau - t, t)$ is Lipschitzian on $\mathbb{R}^n$ with the constant $M + 1$. Pick any u, $u_1 \in \mathbb{R}^n$, $v \in \text{cl}\,\hat{r}(\tau - t, t)u$ and consider $v_i \to v$, $v_i \in \hat{r}(\tau - t, t)u$. By the Lipschitz continuity of $\hat{r}(\tau - t, t)$ for some $w_i \in \hat{r}(\tau - t, t)u_1$, $\|w_i - v_i\| \leqq (M + 1)\|u - u_1\|$. Taking a subsequence and keeping the same notation, we may assume that w_i converges to some $w \in \text{cl}\,\hat{r}(\tau - t, t)u_1$. Then $\|w - v\| \leqq (M + 1)\|u - u_1\|$ and this yields the Lipschitz continuity of cl $\hat{r}(\tau - t, t)$. Let $(u_i, v_i) \in \text{graph}\,(\hat{r}(\tau - t, t))$ be a sequence converging to some (u, v). Then $v_i \in \hat{r}(\tau - t, t)u_i$ and, by Lipschitz continuity, for some $w_i \in \hat{r}(\tau - t, t)u$ we have $\|w_i - v_i\| \leqq (M + 1)\|u - u_i\|$. Hence $w_i \to v$ and $v \in \text{cl}\,\hat{r}(\tau - t, t)u$. This implies that

$$\overline{\text{graph}\,(\hat{r}(\tau - t, t))} = \text{graph}\,(\text{cl}\,\hat{r}(\tau - t, t)) \tag{3.14}$$

and therefore graph $(\operatorname{cl} \hat{r}(\tau-t, t))$ is a closed convex cone. Hence $\operatorname{cl} \hat{r}(\tau-t, t)$ is a closed convex process and

$$\operatorname{graph}(\hat{r}(\tau-t, t))^{+}=\operatorname{graph}(\operatorname{cl} \hat{r}(\tau-t, t))^{+}.$$

From Definition 3.2 we deduce that $\hat{r}(\tau-t, t)^{*}=(\operatorname{cl} \hat{r}(\tau-t, t))^{*}$. The last statements follow from Proposition 1.7 of [3].

Proof of Lemma 3.5. We prove first that

$$\hat{r}_{Q}(\tau-t, t)^{*}=(\operatorname{cl} \hat{r}_{\bar{Q}}(\tau-t, t))^{*}. \tag{3.15}$$

Indeed, fix $u_i \in Q$, $v_i \in \hat{r}(\tau-t, t)u_i$ such that $\lim_{i\to\infty}(u_i, v_i)=(u, v)$. Then $u \in \bar{Q}$ and $(u, v) \in \overline{\operatorname{graph}(\hat{r}(\tau-t, t))}=\operatorname{graph}(\operatorname{cl} \hat{r}(\tau-t, t))$ (by (3.14)). Hence $v \in \operatorname{cl} \hat{r}_{\bar{Q}}(\tau-t, t)u$ and we proved that $\overline{\operatorname{graph}(\hat{r}_{Q}(\tau-t, t))}=\operatorname{graph}(\operatorname{cl} \hat{r}_{\bar{Q}}(\tau-t, t))$ this yields (3.15). We also know that $\operatorname{Dom}(\operatorname{cl} \hat{r}(\tau-t, t))=\mathbb{R}^{n}$. Hence using [2, pp. 142-143] we obtain (3.8).

To prove the second statement we observe that the Lipschitz continuity of $\operatorname{cl} \hat{r}(\tau-t, t)$ yields

$$\overline{\operatorname{cl} \hat{r}(\tau-t, t)\bar{Q}}=\overline{\operatorname{cl} \hat{r}(\tau-t, t)Q}.$$

hence $(\hat{r}(\tau-t, t)Q)^{+}=(\operatorname{cl} \hat{r}(\tau-t, t)Q)^{+}=(\operatorname{cl} \hat{r}(\tau-t, t)\bar{Q})^{+}=$(by [2, pp. 142-143]) $\operatorname{cl} \hat{r}(\tau-t, t)^{*-1}(Q^{+})=$(by Lemma 3.6) $\hat{r}(\tau-t, t)^{*-1}(Q^{+})$. The proof is complete.

4. The cone $T_{R(\tau,C)}(z(\tau))^{+}$. In this section we assume that (H_1) holds true and that there exists a family of closed convex processes $\{A(s)\}_{s\in[0,T]}$ satisfying (H_2) and (H_3).

Observe that the dual form of Theorem 2.4 is as follows: for all $0 \leqq t<\tau \leqq T$,

$$T_{R(\tau,C)}(z(\tau))^{+} \subset (r(\tau-t, t)W(t, z))^{+}. \tag{4.1}$$

Hence we can "estimate" $T_{R(\tau,C)}(z(\tau))^{+}$ using the set $(r(\tau-t, t)W(t, z))^{+}$. We study this last set via a duality technique.

Consider again the adjoint differential inclusion

$$-q'(s) \in A(s)^{*}q(s) \quad \text{a.e.} \tag{4.2}$$

THEOREM 4.1. *Assume that* (H_1)-(H_3) *hold true. Let* $Q(t) \subset W(t, z)$ *be a family of convex cones such that for all* $0 \leqq t \leqq t_1 \leqq T$, $\hat{r}(t_1-t, t)Q(t) \subset Q(t_1)$. *Then for all* $\tau \in [0, T]$

$$T_{R(\tau,C)}(z(\tau))^{+} \subset \{q(\tau): q \in W^{1,\infty}(0, \tau) \text{ satisfies (4.2)}, q(t) \in Q(t)^{+} \text{ on } [0, \tau[\}.$$

Consider next the differential inclusion

$$\begin{cases} -q'(s) \in A(s)^{*}q(s) & \text{a.e.,} \\ \langle q(s), z'(s)\rangle=\min\{\langle q(s), e\rangle: e \in F(z(s))\} & \text{a.e.} \end{cases} \tag{4.3}$$

THEOREM 4.2. *Assume that* (H_1)-(H_3) *hold true and let* $Q(t) \subset \mathcal{W}(t, z)$ *be any family of convex cones. Then for all* $\tau \in [0, T]$

$$T_{R(\tau,C)}(z(\tau))^{+} \subset \{q(\tau): q \in W^{1,\infty}(0, \tau) \text{ satisfies (4.3)}, q(t) \in Q(t)^{+} \text{ on } [0, \tau[\}.$$

In particular,

$$T_{R(\tau,C)}(z(\tau))^{+} \subset \{q(\tau): q \in W^{1,\infty}(0, \tau) \text{ satisfies (4.3)}, q(t) \in R^{\infty}(t, z)^{+}\}.$$

Observe that the statements of the above theorems depend on the choice of $\{A(s)\}$ and $\{Q(s)\}$. From (4.1) and Theorem 3.3(c) we obtain Lemma 4.3.

LEMMA 4.3. *If* (H_1)-(H_3) *hold true, then for any* $0 \leqq t < \tau \leqq T$ *and any convex cone* $Q \subset W(t, z)$

$$T_{R(\tau,C)}(z(\tau))^+ \subset \{q(\tau): q \in W^{1,\infty}(t, \tau) \text{ satisfies (4.2)}, q(t) \in Q^+\}.$$

Proof of Theorem 4.1. We apply the above lemma. Fix $\tau \in]0, T]$ and $b \in T_{R(\tau,C)} \times (z(\tau))^+$.

Step 1. Fix any $0 \leqq t_1 < \cdots < t_m < \tau$. We first prove the existence of $q \in W^{1,\infty}(0, \tau)$ satisfying (4.2) such that

$$q(\tau) = b, \tag{4.4}$$

$$q(t_i) \in Q(t_i)^+ \quad \forall i = 1, \cdots, m. \tag{4.5}$$

By the assumptions of the theorem, inclusion (4.5) implies that

$$q(t_i) \in (\hat{r}(t_i - t_{i-1}, t_{i-1})Q(t_{i-1}))^+. \tag{4.6}$$

We proceed by the induction. By Lemma 4.3 there exists $q \in W^{1,\infty}(t_m, \tau)$ satisfying (4.4), (4.5) with $i = m$. Assume that we already know that for some $2 \leqq j \leqq m$ there exists $q \in W^{1,\infty}(t_j, \tau)$ such that (4.2), (4.4), (4.5) hold true with $i \geqq j$. From (4.6) we deduce that $q(t_j) \in (\hat{r}(t_j - t_{j-1}, t_{j-1})Q(t_{j-1}))^+$. Applying Lemmas 3.4, 3.5 with $\tau = t_j$, $b = q(t_j)$, and $t = t_{j-1}$ we prove the existence of $\hat{q} \in W^{1,\infty}(t_{j-1}, t_j)$ satisfying (4.2) such that $\hat{q}(t_j) = q(t_j)$, $\hat{q}(t_{j-1}) \in Q(t_{j-1})^+$. Setting

$$q(s) = \begin{cases} q(s) & \text{when } s \in [t_j, \tau], \\ \hat{q}(s) & \text{when } s \in [t_{j-1}, t_j], \end{cases}$$

we end the proof of Step 1. □

Step 2. Let $t_i \in [0, \tau]$, $i = 1, 2, \cdots$ be a dense subset of $[0, \tau]$. Set

$$L = \{(x, y) \in L^2(0, \tau) \times L^2(0, \tau): x(s) \in A(s)^* y(s) \text{ a.e.}\}.$$

Since $A(s)^*$ are closed convex processes, by Mazur's Lemma, L is weakly closed in $L^2(0, \tau) \times L^2(0, \tau)$. By Step 1, for all $j \geqq 1$ there exists $q_j \in W^{1,\infty}(0, \tau)$, satisfying (4.2) and such that $q_j(\tau) = b$ and for all $1 \leqq i \leqq j$

$$q_j(t_i) \in Q(t_i)^+. \tag{4.7}$$

By Proposition 1.6(b) of [3], for all j and almost all $s \in [0, \tau]$, $\|q_j'(s)\| \leqq k(s)\|q_j(s)\|$. This and Gronwall's Lemma imply that $\{q_j\}$ is bounded in $W^{1,2}(0, \tau)$ and, by reflexivity, it has a weak cluster point q. Since L is weakly closed, q satisfies (4.2) and by (4.7), for all i, $q(t_i) \in Q(t_i)^+$. Fix $t \in [0, \tau]$, $w \in Q(t)$ and let $\{t_{i_k}\}$ be a subsequence converging to t from the right. Since $\{A(s)\}$ satisfies (H_2), by the Filippov Theorem there exist $w_k \in \hat{r}(t_{i_k} - t, t)w$ converging to w. Moreover, for all k, $\langle q(t_{i_k}), w_k \rangle \geqq 0$. Therefore, taking the limit, we get $q(t) \in Q(t)^+$ for all $t \in [0, \tau]$. This ends the proof. □

To prove Theorem 4.2 we need two lemmas.

The next lemma shows how a given family $\{A(s)\}$ can be "increased" to a larger family of closed convex processes still satisfying (H_2), (H_3).

LEMMA 4.4. *For all* $s \in [0, T]$ *such that* $z'(s) \in F(z(s))$ *and for all* $x \in \mathbb{R}^n$ *set*

$$G(s)x = \overline{A(s)x + T_{\text{co } F(z(s))}(z'(s))}$$

and set $G(s) = A(s)$ *for all other* s. *Then* $\{G(s)\}_{s \in [0,T]}$ *are closed convex processes satisfying* (H_2), (H_3) *and* $A(s) \subset G(s)$. *Moreover, for almost all* $s \in [0, T]$ *and all* $q \in \mathbb{R}^n$

$$G(s)^* q = \begin{cases} A(s)^* q & \text{when } q \in (F(z(s)) - z'(s))^+, \\ \varnothing & \text{otherwise.} \end{cases} \tag{4.8}$$

Proof. From the definition of $G(s)$, exactly as in the proof of Lemma 3.6, we deduce that $G(s)(\cdot)$ is $k(s)$-Lipschitz on $\mathbb{R}^n$. By Lemma 2.8 of [12] we know that

$\{G(s)\}$ satisfy (H_3). Since $G(s)(\cdot)$ is continuous and has closed images, graph $(G(s))$ is closed. It is also clear that graph $(G(s))$ is a cone. To prove its convexity it is enough to consider only those $s \in [0, T]$ that satisfy $z'(s) \in F(z(s))$. Fix such s and $u, v \in \mathbb{R}^n$. Since $A(s)$ is a convex process and $T_{\text{co } F(z(s))}(z'(s))$ is a convex cone, we obtain

$$A(s)u + T_{\text{co } F(z(s))}(z'(s)) + A(s)v + T_{\text{co } F(z(s))}(z'(s)) \subset A(s)(u+v) + T_{\text{co } F(z(s))}(z'(s)).$$

This yields that

$$G(s)u + G(s)v \subset \overline{A(s)(u+v) + T_{\text{co } F(z(s))}(z'(s))} = G(s)(u+v).$$

Hence $G(s)$ is a closed convex process. Moreover, by [25], for all $q \in \mathbb{R}^n$

$$G(s)^*q = \begin{cases} A(s)^*q & \text{when } q \in T_{\text{co } F(z(s))}(z'(s))^+, \\ \varnothing & \text{otherwise.} \end{cases} \tag{4.9}$$

Since co $F(z(s))$ is a convex set we also have

$$T_{\text{co } F(z(s))}(z'(s)) = \overline{\bigcup_{i=1,2,\cdots} i(\text{co } F(z(s)) - z'(s))}, \tag{4.10}$$

and therefore

$$T_{\text{co } F(z(s))}(z'(s))^+ = (\text{co } F(z(s)) - z'(s))^+.$$

Using (4.9), we deduce from the last equality that for almost all $s \in [0, T]$, (4.8) holds true. To end the proof it remains to show that for all $x \in \mathbb{R}^n$, the map $s \to G(s)x$ is measurable. Since the map $s \to F(z(s))$ is continuous it is also measurable. By Castaing's Representation Theorem [8] and the assumption (H_2)(i) there exist measurable selections

$$f_n(s) \in F(z(s)), \quad g_n(s) \in A(s)x, \quad n = 1, 2, \cdots,$$

such that for all s

$$\overline{\bigcup_{n \geq 1} f_n(s)} = F(z(s)), \qquad \overline{\bigcup_{n \geq 1} g_n(s)} = A(s)x.$$

Hence, using (4.10), we obtain

$$G(s)x = \overline{\bigcup_{n \geq 1} g_n(s) + \bigcup_{i \geq 1} i \bigcup_{n \geq 1} (f_n(s) - z'(s))} = \overline{\bigcup_{\substack{n \geq 1 \\ i,j \geq 1}} g_n(s) + i(f_j(s) - z'(s))}.$$

Since the functions $s \to g_n(s) + i(f_j(s) - z'(s))$ are measurable, the last equality and Castaing's Theorem imply that $s \to G(s)x$ is a measurable set-valued map. □

In Theorem 4.1 we deal with convex cones $Q(t) \subset W(t, z)$ that have the following invariance property:

$$\forall 0 \leqq t < t_1 \leqq T \quad \hat{r}(t_1 - t, t)Q(t) \subset Q(t_1). \tag{4.11}$$

The next result shows how such cones can be constructed.

LEMMA 4.5. *Let $\{A(s)\}_{s \in [0,T]}$ be any family of closed convex processes satisfying (H_2), (H_3) and $\hat{Q}(t) \subset \mathcal{W}(t, z)$ be convex cones. Then there exist convex cones $Q(t) \supset \hat{Q}(t)$ satisfying* (4.11).

Proof. For all $0 \leqq t_1 \leqq \cdots \leqq t_m \leqq T$, define recursively cones $P(t_1) = \hat{Q}(t_1) + \hat{r}(t_1, 0)\hat{Q}(0), \cdots, P(t_1, \cdots, t_{i+1}) = \hat{Q}(t_{i+1}) + \hat{r}(t_{i+1} - t_i, t_i)P(t_1, \cdots, t_i)$. Using an induction argument, we prove by Theorems 2.4 and 2.7 that for all $i \geqq 1$, $P(t_1, \cdots, t_i) \subset \mathcal{W}(t, z)$. Set

$$Q(t) = \bigcup_{\substack{0 \leqq t_1 \leqq \cdots \leqq t_m = t \\ m \geqq 1}} P(t_1, \cdots, t_m).$$

Clearly $Q(t)$ is a cone containing $\hat{Q}(t)$ and, by definition of $Q(t)$, for all $0 \leqq t \leqq t_1 \leqq T$, $\hat{r}(t_1 - t, t)Q(t) \subset Q(t_1)$. It remains to prove that $Q(t)$ is convex, i.e., we must check that for all $0 \leqq t_1 \leqq \cdots \leqq t_m = t$, $0 \leqq t'_1 \leqq \cdots \leqq t'_k = t$

$$P(t_1, \cdots, t_m) + P(t'_1, \cdots, t'_k) \subset Q(t). \tag{4.12}$$

We proceed by the induction with respect to $m + k$. Observe that for all $t \in [0, T]$, $P(t)$ is a convex cone. Fix $t \in [0, T]$. Assume that for some $j \geqq 2$ and all $m \geqq 1$, $k \geqq 1$, $0 \leqq t_1 \leqq \cdots \leqq t_m = t$, $0 \leqq t'_1 \leqq \cdots \leqq t'_k = t$ satisfying $m + k \leqq j$ the relation (4.12) holds true. Fix $0 \leqq t_1 \leqq \cdots \leqq t_{m+1} = t$, $0 \leqq t'_1 \leqq \cdots \leqq t'_k = t$ such that $m + k = j$, $t_{k-1} \leqq t_m$. Then $P(t_1, \cdots, t_m) + P(t'_1, \cdots, t'_{k-1}, t_m) \subset Q(t_m)$. Moreover, by definition of $P(\cdot)$, using that $\hat{r}$ is a convex process we obtain

$$\hat{r}(t_m - t'_{k-1}, t'_{k-1})P(t'_1, \cdots, t'_{k-1}) \subset P(t'_1, \cdots, t'_{k-1}, t_m).$$

This and the definition of $Q(t)$ imply:

$$\begin{aligned} P(t_1, \cdots, t_{m+1}) + P(t'_1, \cdots, t'_k) &= \hat{Q}(t) + \hat{r}(t - t_m, t_m)P(t_1, \cdots, t_m) \\ &\quad + \hat{Q}(t) + \hat{r}(t - t_m, t_m)\hat{r}(t_m - t'_{k-1}, t'_{k-1})P(t'_1, \cdots, t'_{k-1}) \\ &\subset \hat{Q}(t) + \hat{r}(t - t_m, t_m)(P(t_1, \cdots, t_m) + P(t_1, \cdots, t'_{k-1}, t_m)) \\ &\subset \hat{Q}(t) + \hat{r}(t - t_m, t_m)Q(t_m) \subset Q(t). \end{aligned}$$

Proof of Theorem 4.2. By Lemma 4.4 we replace the family $\{A(s)\}$ by the new family $\{G(s)\}$ satisfying (H_2), (H_3) and (4.8). From Lemma 4.5 it is not restrictive to assume that the family $\{Q(s)\}$ satisfies (4.11). Theorem 4.1 applied with $\{G(s)\}$ yields the result. □

COROLLARY 4.6. *Assume that* (H_1)-(H_3) *hold true and let* Q *be a convex subcone of* $T_C(z(0))$. *Then for all* $\tau \in [0, T]$

$$T_{R(\tau,C)}(z(\tau))^+ \subset \{q(\tau)\colon q \in W^{1,\infty}(0, \tau) \text{ satisfies (4.3)}, q(0) \in Q^+\}.$$

Proof. Setting $Q(t) = \hat{r}(t, 0)Q$ and applying Theorem 4.1 with closed convex processes $\{G(s)\}$ of Lemma 4.4, we deduce our statement from (4.8). □

THEOREM 4.7. *Assume that* (H_1)-(H_3) *hold true and that, for any* $t \in [0, T]$, q_1, $q_2 \in W^{1,\infty}(0, t)$ *satisfying* (4.3) *and equal at* t, *we have* $q_1/\|q_1\| = q_2/\|q_2\|$ *on* $[0, t]$. *Then for all* $\tau \in [0, T]$

$$T_{R(\tau,C)}(z(\tau))^+ \subset \{q(\tau)\colon q \in W^{1,\infty}(0, \tau) \text{ satisfies (4.3) and } q(t) \in W(t, z)^+ \text{ on } [0, \tau[\}.$$

In particular, the above happens when for almost all $s \in [0, T]$, *the adjoint* $A(s)^*$ *is single-valued and Lipschitzian on its domain of definition.*

Proof. Fix $\tau \in [0, T]$, $b \in T_{R(\tau,C)}(z(\tau))^+$, $t \in [0, \tau[$, $c \in W(t, z)$. By Theorem 4.1 applied with the family of closed convex processes $\{G(s)\}$ and the convex cones

$$Q(s) = \begin{cases} \varnothing & \text{for } s < t, \\ \mathbb{R}_+ c & \text{for } s = t, \\ \hat{r}(s - t, t)Q(t) & \text{for } s > t, \end{cases}$$

using (4.8), we prove the existence of $q \in W^{1,\infty}(0, \tau)$ satisfying (4.3) such that $q(\tau) = b$, $\langle q(t), c\rangle \geqq 0$. Since $c \in W(t, z)$ and $t \in [0, \tau[$ are arbitrary, by the assumptions of Theorem 4.1, $q(t) \in W(t, z)^+$ on $[0, \tau[$. □

COROLLARY 4.8. *Assume that* (H_1) *holds true and that there exist linear operators* $A(s) \in L(\mathbb{R}^n, \mathbb{R}^n)$ *satisfying* (H_2), (H_3). *Then for all* $\tau \in (0, T)$,

$$\begin{aligned} T_{R(\tau,C)}(z(\tau))^+ \subset \{q(\tau)\colon & -q'(s) = A(s)^* q(s), \\ & \langle q(s), z'(s)\rangle = \min_{e \in F(z(s))} \langle q(s), e\rangle, \ q(s) \in W(s, z)^+ \text{ in } [0, \tau[\}. \end{aligned}$$

Proof. The transposed linear operator $A(s)^*$ is equal to the adjoint process in the sense of Definition 3.1 (see Rockafellar [25]). Since for all $b \in T_{R(\tau,C)}(z(\tau))^+$, the solution of the linear equation $-q'(s) = A(s)^*q(s)$; $q(\tau) = b$ is unique, the proof follows from Theorem 4.7. □

THEOREM 4.9. *Let $R_C(T, \cdot)$ denote the restriction of the reachable map $R(T, \cdot)$ to the set C. Then for every convex cone $Q \subset T_C(z(0))$,*

$$T_{\text{graph } R_C(T,\cdot)}(z(0), z(T))^+ \subset \{(\pi - q(0), q(T)) : q \in W^{1,\infty}(0, T) \text{ satisfies (4.3) and } \pi \in Q^+\}.$$

Proof. By Theorem 2.6,

$$T_{\text{graph } R_C(T,\cdot)}(z(0), z(T))^+ \subset \{(w(0), r(T, 0)w(0)) : w(0) \in T_C(z(0))\}^+. \tag{4.13}$$

We replace closed convex processes $\{A(s)\}$ by $\{G(s)\}$ from Lemma 4.4 and keep the same notation $\hat{r}$ for the reachable map of the inclusion

$$w'(s) \in G(s)w(s) \quad \text{a.e.}$$

Then by (3.4), (4.13) we obtain

$$T_{\text{graph } R_C(T,\cdot)}(z(0), z(T))^+ \subset \{(a, \hat{r}(T, 0)a) : a \in Q\}^+,$$

and from Lemma 3.5 we deduce that for all $(p, q) \in T_{\text{graph } R_C(T,\cdot)}(z(0), z(T))^+$ we have $p + \hat{r}(T, 0)^*q \in Q^+$. Lemma 3.4 ends the proof. □

Remark 4.10 (*On the Hamiltonian inclusions*). For all $x, p \in \mathbb{R}^n$ the Hamiltonian of F is defined by

$$H(x, p) = \sup_{e \in F(x)} \langle p, e \rangle = \sup_{e \in \text{co } F(x)} \langle p, e \rangle.$$

If (H_1) holds true, then H is locally Lipschitz on Dom $F \times \mathbb{R}^n$ (see, for example, [9]). Let us assume that for all s, Dom $A(s)^*$ is a subspace of $\mathbb{R}^n$ and $A(s)^*$ is *linear on* Dom $A(s)^*$.

Consider an absolutely continuous solution q of (4.3) defined on the time interval $[0, T]$. Pick any $s \in]0, 1[$ such that $\langle q(s), z'(s) \rangle = \min_{e \in F(z(s))} \langle q(s), e \rangle$, $-q'(s) = A(s)^*q(s)$. Set $\bar{q} = -q$ and fix any u. Let $v \in A(s)u$ and $v_h \to v$ (when $h \to 0+$) be such that $z'(s) + hv_h \in \text{co } F(z(s) + hu)$. Then for all $w \in \mathbb{R}^n$ we have

$$\begin{aligned}
\limsup_{h \to 0+} & \frac{H(z(s) + hu, \bar{q}(s) + hw) - H(z(s), \bar{q}(s))}{h} \\
& \geqq \limsup_{h \to 0+} \frac{\langle \bar{q}(s) + hw, z'(s) + hv_h \rangle - \langle \bar{q}(s), z'(s) \rangle}{h} \\
& = \langle w, z'(s) \rangle + \langle \bar{q}(s), v \rangle \\
& \geqq \langle w, z'(s) \rangle + \langle q'(s), u \rangle = \langle (q'(s), z'(s)), (u, w) \rangle.
\end{aligned}$$

In particular this yields

$$(-\bar{q}'(s), z'(s)) \in \partial H(z(s), \bar{q}(s)) \tag{4.14}$$

where ∂H denotes the generalized gradient of H (see [9]). Hence in this particular case for every solution q of (4.3), $-q$ is a solution of the Hamiltonian inclusion (4.14). It may happen that for a family of closed convex processes satisfying (H_2), (H_3), the only solution of (4.3) is $q \equiv 0$, and at the same time the Hamiltonian inclusion (4.14) has solutions different from zero (see the example from [18]). Hence in this particular case it is more convenient to use the adjoint inclusion (4.3) than the Hamiltonian

inclusion (4.14) to estimate the cone $T_{R(T,C)}(z(T))^+$. In a more general case we do not know how to compare solutions of (4.3) and (4.14).

5. Application: high-order maximum principles.

5.1. Minimization with respect to the final state. Let U be a compact metric space and $f:\mathbb{R}^n \times U \to \mathbb{R}^n$ be a continuous function, $g:\mathbb{R}^n \to \mathbb{R}$, $C \subset \mathbb{R}^n$. Consider the following optimal control problem:

$$\text{minimize } g(x(1)) \tag{5.1}$$

over the solutions of the control system

$$\begin{cases} x'(t) = f(x(t), u(t)) & \text{a.e. in } [0,1], \\ x(0) \in C, \qquad u(t) \in U \text{ is measurable.} \end{cases} \tag{5.2}$$

Set $F(x) = f(x, U)$ for all $x \in \mathbb{R}^n$. By the Filippov Theorem [1, p. 91] solutions of the control system (5.2) and the differential inclusion

$$\begin{aligned} &x'(t) \in F(x(t)) \quad \text{a.e. on } [0,1], \\ &x(0) \in C \end{aligned} \tag{5.3}$$

do coincide.

THEOREM 5.1. *Assume that a trajectory control pair $(z, \bar{u})$ solves the above problem and for a constant L and all $u \in U$, $f(\cdot, u)$ is L-Lipschitzian on a neighborhood of $z([0,1])$. If g is differentiable at $z(1)$ and for almost all t, $f(\cdot, u(t))$ is differentiable at $z(t)$, then there exists $q \in W^{1,\infty}(0,1)$ such that*

$$\begin{aligned} &-q'(t) = q(t)\frac{\partial f}{\partial x}(z(t), \bar{u}(t)) \quad \text{a.e.}, \\ &\langle q(t), z'(t)\rangle = \min_{u \in U} \langle q(t), f(z(t), u)\rangle \quad \text{a.e.}, \end{aligned} \tag{5.4}$$

$$q(1) = g'(z(1)), \qquad q(0) \in T_C(z(0))^+, \tag{5.5}$$

$$q(t) \in W(t, z)^+ \quad \text{for all } t \in [0, 1[. \tag{5.6}$$

Proof. By the assumptions, the set-valued map F defined above satisfies (H_1). Moreover, for almost all $s \in [0,1]$, $(\partial/\partial x)f(z(s), \bar{u}(s)) \subset dF(z(s), z'(s)) \subset d \text{ co } F(z(s), z'(s))$. Set $A(s) = (\partial/\partial x)f(z(s), \bar{u}(s))$. Since $\|A(s)\| \leqq L$, then $A(s)$ is L-Lipschitz. Hence (H_2), (H_3) hold true. On the other hand, for every solution x of (5.3) we have $g(x(1)) - g(z(1)) \geqq 0$, which yields

$$\forall w \in T_{R(1,C)}(z(1)) \quad g'(z(1))w \geqq 0,$$

i.e.,

$$g'(z(1)) \in T_{R(1,C)}(z(1))^+.$$

Corollary 4.8 ends the proof. □

COROLLARY 5.2. *Under all assumptions of Theorem 5.1, assume that for some $t \in [0, 1[$, $W(t, z)^+ = \{0\}$. Then $z(1)$ is a critical point of g and if g is locally C^2 at $z(1)$ then $g''(z(1)) \geqq 0$ on $T_{R(1,C)}(z(1))$. In particular, this happens when $T_C(z(0))^+ = \{0\}$.*

Proof. Let q be as in Theorem 5.1 and let t be such that $W(t, z)^+ = \{0\}$. Then $q(t) = 0$ and, by the uniqueness of q, $q(1) = 0$. Hence, by (5.5), $g'(z(1)) = 0$. Assume next that g is locally C^2 and fix $w \in T_{R(1,C)}(z(1))$. Then for some $h_i \to 0+$, $w_i \to w$, $z(1) + h_i w_i \in R(1, C)$ and since z solves the problem (5.1), (5.2), $g(z(1) + h_i w_i) - g(z(1)) = \frac{1}{2}g''(z(1))w_i w_i h_i^2 + o(h_i^2) \geqq 0$. Taking the limit, we end the proof. □

5.2. Minimization with respect to both endpoints. Let f, U be as in § 5.1 and $\varphi : \mathbb{R}^{2n} \to \mathbb{R}$ be a given function. Consider the problem

$$\text{minimize } \varphi(x(0), x(1)) \tag{5.7}$$

over the solutions of the control system (5.2). If a trajectory-control pair $(z, \bar{u})$ solves the problem (5.7), (5.2) and φ is differentiable at $(z(0), z(1))$, then

$$\forall (w(0), w(1)) \in T_{\text{graph } R_C(1,\cdot)}(z(0), z(1)) \quad \varphi'(z(0), z(1))(w(0), w(1)) \geqq 0,$$

i.e., $\varphi'(z(0), z(1))$ is in the positive polar of the tangent cone. Let $W(t, z)$ denote the cone of variations of reachable sets $R(\cdot, z(0))$.

THEOREM 5.3. *Assume that a trajectory-control pair $(z, \bar{u})$ solves the above problem, f satisfies all the assumptions of Theorem* 5.1, *and φ is differentiable at $(z(0), z(1))$. Then there exists $q \in W^{1,\infty}(0, 1)$ satisfying* (5.4), (5.6) *and such that*

$$q(1) = \frac{\partial}{\partial x_2} \varphi(z(0), z(1)), \qquad q(0) \in T_C(z(0))^+ - \frac{\partial}{\partial x_1} \varphi(z(0), z(1)).$$

Proof. By the proof of Theorem 5.1 the family of maps $A(s) = (\partial/\partial x) f(z(s), \bar{u}(s))$, $s \in [0, 1]$ satisfies (H_2), (H_3). We already know that $\varphi'(z(0), z(1)) \in T_{\text{graph } R_C(1,\cdot)}(z(0), z(1))^+$. Fix $b \in T_C(z(0))$. Applying Theorem 4.9 with $Q = \mathbb{R}_+ b$ we deduce that the solution q of (5.4) satisfying $q(1) = (\partial/\partial x_2)\varphi(z(0), z(1))$ verifies

$$\frac{\partial}{\partial x_1} \varphi(z(0), z(1)) \in (\mathbb{R}_+ b)^+ - q(0).$$

Hence $\langle q(0) + (\partial/\partial x_1)\varphi(z(0), z(1)), b \rangle \geqq 0$. Since q does not depend on b, we obtain that $q(0) + (\partial/\partial x_1)\varphi(z(0), z(1)) \in T_C(z(0))^+$. It remains to show that q satisfies (5.6). Set $g(x) = \varphi(z(0), x)$. Then $g'(z(1)) = (\partial/\partial x_2)\varphi(z(0), z(1))$. Clearly, $(z, \bar{u})$ is an optimal solution of problem (5.1), (5.2) with $C = \{z(0)\}$. Applying Theorem 5.1 with $C = \{z(0)\}$ we end the proof. □

COROLLARY 5.4. *Under all assumptions of Theorem* 5.3 *assume that for some $t \in [0, 1[$, $W(t, z)^+ = \{0\}$. Then $(\partial/\partial x_1)\varphi(z(0), z(1)) \in T_C(z(0))^+$, $(\partial\varphi/\partial x_2)(z(0), z(1)) = 0$. Moreover, if $T_C(z(0))^+ = \{0\}$, then $(z(0), z(1))$ is a critical point of φ, and if φ is locally C^2 at $(z(0), z(1))$, then $\varphi''(z(0), z(1)) \geqq 0$ on $T_{\text{graph } R_C(1,\cdot)}(z(0), z(1))$.*

The proof follows by the same arguments as in Corollary 5.2.

5.3. Closed-loop control systems. Let $U : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued map with compact nonempty images, let C be a nonempty subset of $\mathbb{R}^n$, and let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitzian function, $g : \mathbb{R}^n \to \mathbb{R}$. Consider the following control problem:

$$\text{minimize } g(x(1)) \tag{5.8}$$

over trajectories of the control system

$$\begin{aligned} &x'(t) = f(x(t), u(t)) \quad \text{a.e. in } [0, 1], \\ &x(0) \in C, \qquad u(t) \in U(x(t)) \text{ is measurable.} \end{aligned} \tag{5.9}$$

Set $F(x) = \{f(x, u) : u \in U(x)\}$. It is clear that every trajectory of (5.9) is a trajectory of the differential inclusion

$$\begin{aligned} &x'(t) \in F(x(t)) \quad \text{a.e. in } [0, 1], \\ &x(0) \in C. \end{aligned} \tag{5.10}$$

LEMMA 5.5. *If U is upper semicontinuous, then the set of trajectories of the closed-loop control system* (5.9) *do coincide with the set of trajectories of the differential inclusion* (5.10).

Proof. We must show that, with every trajectory $x \in W^{1,1}(0,1)$ of the inclusion (5.10), we can associate a measurable function $u\colon [0,1] \to \mathbb{R}^m$ satisfying

$$x'(t) = f(x(t), u(t)), \quad u(t) \in U(x(t)) \quad \text{a.e. in } [0,1].$$

For all $t \in [0,1]$ set $\hat{U}(t) = \{u \in U(x(t))\colon x'(t) = f(x(t), u)\}$. Then for almost all $t \in [0,1]$, $\hat{U}(t)$ is a closed, nonempty set. We claim that $\hat{U}$ is a measurable set-valued map. Indeed fix a closed subset $\hat{C} \subset \mathbb{R}^m$ and observe that the set

$$D := \{(t, f(x(t), u))\colon t \in [0,1], u \in U(x(t)) \cap \hat{C}\}$$

is closed. Moreover,

$$\{t\colon \hat{U}(t) \cap \hat{C} \neq \varnothing\} = \{t\colon (t, x'(t)) \in D\}.$$

Thus $\{t\colon \hat{U}(t) \cap \hat{C} \neq \varnothing\}$ is a Lebesgue measurable set and, since $\hat{C}$ is an arbitrary closed subset of $\mathbb{R}^m$, we proved that $\hat{U}$ is measurable. From the measurable selection theorem (see, for example, [26]) follows the existence of a measurable selection $u(t) \in \hat{U}(x(t))$, $t \in [0,1]$. The very definition of the map $\hat{U}$ ends the proof. □

In the theorem below we assume that $f(x, U(x))$ is *regular* in the following sense: If for some x and $\bar{u} \in U(x)$, $q \neq q_1 \neq 0$ we have

$$\sup_{u \in U(x)} \langle q, f(x,u)\rangle = \langle q, f(x,\bar{u})\rangle, \qquad \sup_{u \in U(x)} \langle q_1, f(x,u)\rangle = \langle q_1, f(x,\bar{u})\rangle,$$

then for some $\lambda \geqq 0$ $q = \lambda q_1$. Geometrically this means that every boundary point of $\operatorname{co} f(x, U(x))$ has at most one normalized outer normal.

THEOREM 5.6. *Assume that a trajectory control pair $(z, \bar{u})$ solves the problem above, and that f is differentiable at $(z(t), \bar{u}(t))$, g is differentiable at $z(1)$, U is Lipschitzian on a neighborhood of $z([0,1])$, and $f(x, U(x))$ is regular. Further, assume that there exist closed convex processes $B(s) \subset dU(z(s), \bar{u}(s))$ satisfying (H_2). Then there exists a solution $q \in W^{1,\infty}(0,1)$ of the inclusion*

$$-q' \in \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* q + B(t)^* \frac{\partial f}{\partial u}(z(t), \bar{u}(t))^* q$$

satisfying (5.5), (5.6) *and the minimum principle*

$$\langle q(t), z'(t)\rangle = \min_{u \in U(z(t))} \langle q(t), f(z(t), u)\rangle \quad \text{a.e.}$$

Proof. From differentiability of f at $(z(t), \bar{u}(t))$ we deduce that for almost all t and for all $w \in \mathbb{R}^n$

$$\frac{\partial f}{\partial x}(z(t), \bar{u}(t))w + \frac{\partial f}{\partial u}(z(t), \bar{u}(t))\, dU(z(t), \bar{u}(t))w \subset dF(z(t), z'(t))w.$$

Hence the closed convex processes

$$A(t) := \frac{\partial f}{\partial x}(z(t), \bar{u}(t)) + \frac{\partial f}{\partial u}(z(t), \bar{u}(t))B(t)$$

satisfy (H_2), (H_3). Since z is the minimizing trajectory for all $w \in T_{R(1,C)}(z(1))$, $g'(z(1))w \geqq 0$. Thus, $g'(z(1)) \in T_{R(1,C)}(z(1))^+$. We apply Theorem 4.7. Let q_1, q_2 be two solutions of (4.3) such that $q_1(t) = q_2(t) \neq 0$. Then $q_i \neq 0$ on $[0,t]$ and

$$\langle q_1(s), z'(s)\rangle = \min_{e \in F(z(s))} \langle q_1(s), e\rangle \quad \text{a.e.},$$

$$\langle q_2(s), z'(s)\rangle = \min_{e \in F(z(s))} \langle q_2(s), e\rangle \quad \text{a.e.}$$

Since $f(x, U(x))$ is regular $q_1(s)/\|q_1(s)\| = q_2(s)/\|q_2(s)\|$ a.e. in $[0, 1]$ and, by continuity of $q(\cdot)$ we obtain $q_1/\|q_1\| = q_2/\|q_2\|$. Hence the result will follow from Theorem 4.7 if we show that

$$A(t)^* \subset \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* + B(t)^* \frac{\partial f}{\partial u}(z(t), \bar{u}(t))^*.$$

Fix $p \in A(t)^* q$. Then for all $w \in \mathbb{R}^n$, $v \in B(t)w$

$$\begin{aligned}\langle p, w\rangle &\leqq \langle q, \partial f/\partial x(z(t), \bar{u}(t))w + \partial f/\partial u\,(z(t), \bar{u}(t))v\rangle \\ &= \langle \partial f/\partial x(z(t), \bar{u}(t))^* q, w\rangle + \langle \partial f/\partial u(z(t), \bar{u}(t))^* q, v\rangle\end{aligned}$$

and therefore

$$\langle p - \partial f/\partial x(z(t), \bar{u}(t))^* q, w\rangle \leqq \langle \partial f/\partial u(z(t), \bar{u}(t))^* q, v\rangle.$$

By the definition of the adjoint process

$$p - \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* q \in B(t)^* \frac{\partial f}{\partial u}(z(t), \bar{u}(t))^* q,$$

and we finally obtain

$$p \in \frac{\partial f}{\partial x}(z(t), \bar{u}(t))^* q + B(t)^* \frac{\partial f}{\partial u}(z(t), \bar{u}(t))^* q.$$

The proof is complete. □

The next result is an extension of the main theorem from [22].

THEOREM 5.7. *Assume that a trajectory control pair $(z, \bar{u})$ solves the problem* (5.8), (5.9) *and that f is differentiable at $(z(t), \bar{u}(t))$, g is differentiable at $z(1)$, and U is Lipschitzian on a neighborhood of $z([0, 1])$. Further, assume that for almost all t there exists a selection $u_t(x) \in U(x)$ that is differentiable at $z(t)$ and satisfies $u_t(z(t)) = \bar{u}(t)$. Then there exists a solution $q \in W^{1,\infty}(0, 1)$ of the equation*

$$\begin{aligned} &-q' = q\left(\frac{\partial f}{\partial x}(z(t), \bar{u}(t)) + \frac{\partial f}{\partial u}(z(t), \bar{u}(t))\frac{\partial u_t}{\partial x}(z(t))\right), \\ &\langle q(t), z'(t)\rangle = \min_{u \in U(z(t))} \langle q(t), f(z(t), u)\rangle \text{ a.e.} \end{aligned} \tag{5.11}$$

satisfying (5.5) *and* (5.6).

The above theorem was proved by Leitmann in [22] without the inclusion (5.6).

Proof. The set-valued map $F(x) = f(x, U(x))$ satisfies the hypothesis (H_1) on a neighborhood of $z([0, 1])$. Moreover, the linear operators

$$A(t) = \frac{\partial f}{\partial x}(z(t), u(t)) + \frac{\partial f}{\partial u}(z(t), \bar{u}(t))\frac{\partial u_t}{\partial x}(z(t)), \qquad t \in [0, 1],$$

verify (H_2) and (H_3). Since z is the minimizing trajectory for all $w \in T_{R(1,C)}(z(1))$, $g'(z(1))w \geqq 0$. Thus $g'(z(1)) \in T_{R(1,C)}(z(1))^+$ and the result follows from Corollary 4.8 and the inclusion $W(0, z)^+ \subset T_C(z(0))^+$. □

5.4. An implicit dynamical system. Consider a continuously differentiable function $f: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ and a function $g: \mathbb{R}^n \to \mathbb{R}$, $C \subset \mathbb{R}^n$.

Here we study the problem

$$\text{minimize } g(x(1)) \tag{5.12}$$

over the absolutely continuous solutions of the implicit dynamical system

$$f(x(t), x'(t)) = 0 \quad \text{a.e. in } [0, 1] \tag{5.13}$$

satisfying the initial point constraint

$$x(0) \in C. \tag{5.14}$$

Such systems arise as models for nonlinear circuits. In general they cannot be reduced to the state variable form $z' = f(z, t)$ or to the control system (5.2) (see [6, p. 147, bibliographical comments].

Set $F(x) = \{v: f(x, v) = 0\}$ and consider the differential inclusion

$$x'(t) \in F(x(t)) \quad \text{a.e. in } [0, 1]. \tag{5.15}$$

Clearly solutions of (5.13) and (5.15) do coincide. Moreover, by continuity of f, graph (F) is a closed set. The following result is proved in [15]:

LEMMA 5.8. *Assume that for all $\bar{x} \in \mathbb{R}^n$ there exists $\varepsilon > 0$ such that*

$$\liminf_{\|v\| \to \infty} \inf_{\|x - \bar{x}\| \leqq \varepsilon} \|f(x, v)\| > 0. \tag{5.16}$$

Then F has compact images. If, moreover, for all $(x, v) \in \operatorname{graph}(F)$ the derivative $(\partial/\partial v) f(x, v)$ is surjective, then Dom F *is open and F is locally Lipschitzian on it, and*

$$\ker f'(x, v) = \operatorname{graph}(dF(x, v)).$$

In particular, this implies that $dF(x, v)$ is a closed convex process.

LEMMA 5.9. *Under all assumptions of Lemma 5.8 for every solution x of (5.13) there exist $L > 0$ such that for almost all $s \in [0, 1]$, $dF(x(s), x'(s))$ is L-Lipschitz on $\mathbb{R}^n$ and*

$$dF(x(s), x'(s))^* q = \begin{cases} -\dfrac{\partial f}{\partial x}(x(s), x'(s))^* \dfrac{\partial f}{\partial v}(x(s), x'(s))^{*-1} q & \text{if } q \in \ker \dfrac{\partial f}{\partial v}(x(s), x'(s))^{\perp} \\ \varnothing \quad \text{otherwise.} \end{cases}$$

Proof. Fix a solution x of (5.13). Since the derivative $\partial f/\partial v$ is surjective on graph (F), for all $(x, y) \in \operatorname{graph}(F)$ there exists $\rho > 0$ such that

$$\{v \in \mathbb{R}^m : \|v\| \leqq \rho\} \subset \frac{\partial f}{\partial v}(x, y)(\{u \in \mathbb{R}^n : \|u\| \leqq 1\}). \tag{5.17}$$

Since $f \in C^1$, the assumption (5.16) implies that there exists a compact set K such that for almost all $s \in [0, 1]$, $(x(s), x'(s)) \in K$. This, (5.17) and continuity of $\partial f/\partial v$ imply that for some $\rho > 0$ and almost all $s \in [0, 1]$

$$\{v \in \mathbb{R}^m : \|v\| \leqq \rho\} \subset \frac{\partial f}{\partial v}(x(s), x'(s))(\{u \in \mathbb{R}^n : \|u\| \leqq 1\}).$$

Using Theorem 10.1 of [15] again, we deduce that for some $L > 0$ and almost all $s \in [0, 1]$, $dF(x(s), x'(s))$ is L-Lipschitz on a neighborhood of zero. Since $dF(x(s), x'(s))$ is a convex process, we finally obtain that it is L-Lipschitz on $\mathbb{R}^n$. By the definition of the adjoint process $p \in dF(x(s), x'(s))^* q$ if and only if

$$(p, -q) \in (\ker f'(x(s), x'(s)))^{\perp} = \operatorname{Im} f'(x(s), x'(s))^*.$$

Hence for all $(p, q) \in \operatorname{graph}(dF(x(s), x'(s))^*)$ there exists $\alpha \in \mathbb{R}^m$ such that

$$p = \frac{\partial f}{\partial x}(x(s), x'(s))^* \alpha, \qquad -q = \frac{\partial f}{\partial v}(x(s), x'(s))^* \alpha.$$

Since $(\partial f/\partial v)(x(s), x'(s))$ is surjective, the adjoint linear operator $(\partial f/\partial v)(x(s), x'(s))^*$ is injective and hence invertible on

$$\operatorname{Im} \frac{\partial f}{\partial v}(x(s), x'(s))^* = \left(\ker \frac{\partial f}{\partial v}(x(s), x'(s))\right)^{\perp}.$$

Thus,

$$q \in \left(\ker \frac{\partial f}{\partial v}(x(s), x'(s))\right)^{\perp}, \qquad -p = \frac{\partial f}{\partial x}(x(s), x'(s))^* \frac{\partial f}{\partial v}(x(s), x'(s))^{*-1} q.$$

THEOREM 5.10. *Assume that z solves the problem* (5.12)-(5.14), *f satisfies all the assumptions of Lemma* 5.8, *and g is differentiable at $z(1)$. Then there exists $q \in w^{1,\infty}(0, 1)$ satisfying*

$$q'(s) = \frac{\partial f}{\partial x}(z(s), z'(s))^* \frac{\partial f}{\partial v}(z(s), z'(s))^{*-1} q(s) \quad \text{a.e.}, \tag{5.18}$$

$$q(1) = g'(z(1)), \qquad q(s) \in \left(\ker \frac{\partial f}{\partial v}(x(s), x'(s))\right)^{\perp}, \tag{5.19}$$

$$\min\{\langle q(s), e\rangle : f(z(s), e) = 0\} = \langle q(s), z'(s)\rangle \quad \text{a.e.}, \tag{5.20}$$

$$q(s) \in W(s, z)^+ \quad \text{for } s \in [0, 1[. \tag{5.21}$$

Proof. For all $w \in T_{R(1,C)}(z(1))$, $g'(z(1))w \geqq 0$. Hence $g'(z(1)) \in T_{R(1,C)}(z(1))^+$. Since the solution of (5.18) is uniquely defined, we may apply Theorem 4.7 with closed convex processes $\{dF(x(s), x'(s))\}_{s\in[0,1]}$. Lemma 5.9 ends the proof. □

6. An impulse closed-loop deterministic control problem. Let $U : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ be a set-valued map with compact nonempty images, let C be a nonempty subset of $\mathbb{R}^n$, and let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a locally Lipschitzian function, $g : \mathbb{R}^n \to \mathbb{R}$.

Further, let $V : \mathbb{R}^n \rightrightarrows \mathbb{R}^p$ be a set-valued map of shift parameters and $\varphi : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ be a given function.

Consider the closed-loop control system

$$\begin{aligned} &x'(t) = f(x(t), u(t)), \quad u(t) \in U(x(t)) \quad \text{a.e. in } [0, 1], \\ &\mathrm{x}(0) \in \mathrm{C}. \end{aligned} \tag{6.1}$$

A sequence $\{(t_i, v_i) : i = 1, \cdots, j\}$ is called an impulse strategy of a left-continuous trajectory $x : [0, 1] \to \mathbb{R}^n$, if $0 = t_1 \leqq \cdots \leqq t_j = 1$, and for all i

$$v_i \in V(x(t_i)), \tag{6.2}$$

$$x \in W^{1,1}(t_i, t_{i+1}), \tag{6.3}$$

$$x(t_i+) = x(t_i) + \varphi(x(t_i), v_i), \tag{6.4}$$

and x satisfies (6.1) with a measurable control u. Such trajectory x is called admissible.

This type of system is met in a number of optimal control problems in economics and management (see, for example, [7, pp. 281-285]). We refer to [5], [24], and the references therein for previous results on discontinuous optimal trajectories.

Consider a function $g : \mathbb{R}^n \to \mathbb{R}$. The problem we study here consists in characterization of a solution z to the problem

$$\min\{g(x(1)) : x \text{ is an admissible trajectory}\}. \tag{6.5}$$

The approach is essentially the same. So we shall only stress the main points. For all $x \in \mathbb{R}^n$ set $F(x) = f(x, U(x))$. We prove Lemma 6.1 exactly as we did Lemma 5.5.

LEMMA 6.1. *If U is upper semicontinuous, then the set of admissible trajectories coincide with the set of left-continuous functions $x:[0,1]\to\mathbb{R}^n$ satisfying for some $0=t_1\leqq\cdots\leqq t_j=1$ and some $v_i\in V(x(t_i))$ the following relations*:

$$\begin{aligned} &x\in W^{1,1}(t_i,t_{i+1}),\\ &x'(t)\in F(x(t)) \quad a.e.,\\ &x(0)\in C,\\ &x(t_i+)=x(t_i)+\varphi(x(t_i),v_i). \end{aligned} \tag{6.6}$$

THEOREM 6.2. *Assume that a trajectory-control pair $(z,\bar{u})$ solves the problem above and let $\{(t_i,v_i): i=1,\cdots,l\}$ be a corresponding strategy. Further, assume that U, $\bar{u}$, g, f satisfy all the assumptions of Theorem 5.7, that φ is differentiable at $(z(t_i),v_i)$, and for all i there exists a differentiable at $z(t_i)$ selection $\nu_i(x)\in V(x)$ such that $\nu_i(z(t_i))=v_i$. Then there exists a (left-continuous) function $q:[0,1]\to\mathbb{R}^n$ satisfying* (5.5), (5.11) *and such that for all i*

$$q\in W^{1,\infty}(t_i,t_{i+1}), \tag{6.7}$$

$$q(t_i)=q(t_i+)\left[\mathrm{id}+\frac{\partial\varphi}{\partial x}(z(t_i),v_i)+\frac{\partial\varphi}{\partial v}(z(t_i),v_i)\frac{\partial\nu_i}{\partial x}(z(t_i))\right], \tag{6.8}$$

$$q(t_i+)\in T_{\varphi(z(t_i),V(z(t_i)))}(\varphi(z(t_i),v_i))^+. \tag{6.9}$$

Furthermore, we have the following:

(a) *If the right derivative $z'(t_i+)$ does exist, then*

$$\min_{u\in U(z(t_i))}\langle q(t_i),f(z(t_i),u)\rangle\geqq\langle q(t_i+),z'(t_i+)\rangle;$$

(b) *If the left derivative $z'(t_i-)$ does exist, then*

$$\min_{u\in U(z(t_i+))}\langle q(t_i+),f(z(t_i+),u)\rangle\geqq\langle q(t_i),z'(t_i-)\rangle;$$

(c) *If z has the right and left derivatives at t_i, then*

$$\begin{aligned} \min_{u\in U(z(t_i+))}\langle q(t_i+),f(z(t_i+),u)&=\min_{u\in U(z(t_i))}\langle q(t_i),f(z(t_i),u)\rangle=\langle q(t_i),z'(t_i-)\rangle\\ &=\langle q(t_i+),z'(t_i+)\rangle. \end{aligned} \tag{6.10}$$

When U does not depend on x the assumption that $f(x,\cdot)$ is locally Lipschitzian can be omitted and we have Theorem 6.3.

THEOREM 6.3. *Let U be a compact metric space of controls, V be a set of shift parameters, $f:\mathbb{R}^n\times U\to\mathbb{R}^n$ be a continuous function, and $\varphi:\mathbb{R}^n\times V\to\mathbb{R}^n$. Assume that a trajectory-control pair (z,u) solves the problem*

$$\text{minimize } g(x(1)) \tag{6.11}$$

over the solution of the system

$$\begin{aligned} &x'(t)=f(x(t),u(t)),\quad u(t)\in U \quad a.e.\ in\ [0,1],\\ &x(0)\in C \quad and\ for\ some\ 0=t_1\leqq\cdots\leqq t_j=1,\\ &v_i\in V \quad and\ all\ i,\quad x\in W^{1,1}(t_i,t_{i+1}),\\ &x(t_i+)=x(t_i)+\varphi(x(t_i),v_i), \end{aligned} \tag{6.12}$$

and let $\{(t_i, v_i)\colon i=1, \cdots, l\}$ *be a strategy of* z. *If* f, g *satisfy all the assumptions of Theorem* 5.1 *and* $\varphi(\cdot, v_i)$ *is differentiable at* $z(t_i)$, *then there exists a left-continuous function* $q\colon[0,1]\to\mathbb{R}^n$ *satisfying* (5.4), (5.5), (6.7), (a)-(c) *of Theorem* 6.2 *and*

$$q(t_i) = q(t_i+)\left(\mathrm{id}+\frac{\partial\varphi}{\partial x}(z(t_i), v_i)\right), \tag{6.8'}$$

$$q(t_i+) \in T_{\varphi(z(t_i),V)}(\varphi(z(t_i), v_i))^+. \tag{6.9'}$$

As in § 2 we associate the reachable set $R(t, C)$ at time t with the differential inclusion (6.6).

To prove the Theorem 6.2 we need the following (simple) lemmas.

LEMMA 6.4. *For all* $i=1, \cdots, l-1$ *set*

$$C_i = R(t_i, C) \cup \{x+\varphi(x, v)\colon x\in R(t_i, C), v\in V(x)\}.$$

Then

$$T_{\varphi(z(t_i),V(z(t_i)))}(\varphi(z(t_i), v_i)) \subset T_{C_i}(z(t_i+)).$$

The proof follows from the inclusion $z(t_i)+\varphi(z(t_i), V(z(t_i)))\subset C_i$ and the definition of the contingent cone.

LEMMA 6.5. *For all* $i=1, \cdots, l-1$ *set*

$$A_i = \mathrm{id}+\frac{\partial\varphi}{\partial x}(z(t_i), v_i)+\frac{\partial\varphi}{\partial v}(z(t_i), v_i)\frac{\partial\nu_i}{\partial x}(z(t_i)).$$

Then

$$A_i(T_{R(t_i,C)}(z(t_i))) \subset T_{C_i}(z(t_i+)).$$

Proof. Fix $1\leqq i\leqq l-1$, $w\in T_{R(t_i,C)}(z(t_i))$ and let $h_j\to 0+$, $w_j\to w$ be such that $z(t_i)+h_jw_j\in R(t_i, C)$. Then

$$z(t_i)+h_jw_j+\varphi(z(t_i)+h_jw_j, \nu_i(z(t_i)+h_jw_j))$$

$$= z(t_i+)+h_jw_j+\frac{\partial\varphi}{\partial x}(z(t_i), v_i)h_jw_j+\frac{\partial\varphi}{\partial v}(z(t_i), v_j)\frac{\partial\nu_i}{\partial x}(z(t_i))h_jw_j+o(h_j)\in C_i.$$

The definition of the contingent cone ends the proof. □

LEMMA 6.6. *Assume that* z *has the right derivative* $z'(t_i+)$ *at* t_i *and let* $u\in U(z(t_i))$. *Then the solution* w *of the linear system*

$$w' = \left[\frac{\partial f}{\partial x}(z(t), \bar u(t))+\frac{\partial f}{\partial u}(z(t), \bar u(t))\frac{\partial u_t}{\partial x}(z(t))\right]w,$$

$$w(t_i) = A_if(z(t_i), u)-z'(t_i+)$$

satisfies

$$w(t_{i+1})\in T_{R(t_{i+1},C)}(z(t_{i+1})).$$

Proof. Fix $h_j\to 0+$ and let x be a solution of the inclusion

$$x'(t)\in F(x(t)) \quad \text{a.e. in } [t_i, t_{i+1}],$$

$$x(t_i) = z(t_i), \qquad x'(t_i) = f(z(t_i), u).$$

Then

$$x(t_i+h_j) = z(t_i)+h_jf(z(t_i), u)+o(h_j)\in R(t_i+h_j, C),$$

and therefore

$$x(t_i+h_j)+\varphi(x(t_i+h_j), \nu_i(x(t_i+h_j)))=z(t_i+)+h_jA_if(z(t_i), u)+o(h_j).$$

Thus

$$x(t_i+h_j)+\varphi(x(t_i+h_j), \nu_i(x(t_i+h_j)))=z(t_i+h_j)+h_j[A_if(z(t_i), u)-z'(t_i+)]+o(h_j)$$

and $A_if(z(t_i), u)-z'(t_i+)$ can be seen as a variation of $R(\cdot, C)$ at $(t_i, x(t_i+))$. The proof then follows by the same arguments as Theorem 2.4. □

LEMMA 6.7. *Assume that z has the left derivative $z'(t_i-)$ at t_i. Then for all $u\in U(z(t_i+))$*

$$f(z(t_i+), u)-A_iz'(t_i-)\in T_{C_i}(z(t_i+)).$$

Proof. Fix $h_j\to 0+$, $u\in U(z(t_i+))$ and set

$$x_j:=z(t_i-h_j)+\varphi(z(t_i-h_j), \nu(z(t_i-h_j))).$$

Since F is locally Lipschitzian there exists $M>0$ such that, for all j and $t\in[t_i-h_j, t_i]$, dist $(f(z(t_i+), u)$,

$$F(x_j+(t-t_i+h_j)f(z(t_i+), u))\leqq M(\|x_j-z(t_i+)\|+h_j\|f(z(t_i+), u)\|)=O(h_j).$$

This and Filippov's Theorem imply that

$$x_j+h_jf(z(t_i+), u)\in R(t_i, C)+o(h_j).$$

The definitions of x_j and of the contingent cone end the proof. □

LEMMA 6.8. *For all $p\in T_{R(t_{i+1}, C)}(z(t_{i+1}))^+$ there exists $q\in W^{1,\infty}(t_i, t_{i+1})$ satisfying* (5.11), *such that*

$$q(t_{i+1})=p, \qquad q(t_i+)\in T_{\varphi(z(t_i), V(z(t_i)))}(\varphi(z(t_i), v_i))^+, \tag{6.13}$$

$$q(t_i+)A_i\in T_{R(t_i, C)}(z(t_i))^+. \tag{6.14}$$

Moreover, q satisfies (a)-(c) *of Theorem* 6.2 *with $q(t_i)=q(t_i+)A_i$.*

Proof. Consider the differential inclusion

$$\begin{aligned} &x'(t)\in f(x(t), U(x(t))) \quad \text{a.e. in } [t_i, t_{i+1}],\\ &x(t_i)\in C_i \end{aligned} \tag{6.15}$$

and observe that its reachable set $\hat{R}(t_{i+1}, C_i)$ at time t_{i+1} is contained in $R(t_{i+1}, C)$. Thus $p\in T_{\hat{R}(t_{i+1}, C_i)}(z(t_{i+1}))^+$. By Corollary 4.8 applied on the time interval $[t_i, t_{i+1}]$ to (6.15) and linear operators

$$A(t)=\frac{\partial f}{\partial x}(z(t), \bar{u}(t))+\frac{\partial f}{\partial u}(z(t), \bar{u}(t))\frac{\partial u_t}{\partial x}(z(t)),$$

there exists $q\in W^{1,\infty}(t_i, t_{i+1})$ satisfying (5.11) such that $q(t_{i+1})=p$ and

$$q(t_i+)\in T_{C_i}(z(t_i+))^+. \tag{6.16}$$

Then (6.13) follows from (6.16) and Lemma 6.4 and (6.14) results from (6.16) and Lemma 6.5. Lemma 6.7 and (6.16) imply Theorem 6.2(b). Since q solves the linear equation (5.11), Lemma 6.6 implies that for all $u\in U(z(t_i))$

$$\langle q(t_i), A_if(z(t_i), u)-z'(t_i+)\rangle\geqq 0.$$

Hence we have Theorem 6.2(a). On the other hand, by [13]

$$z'(t_i-)\in\overline{\text{co}}\, F(z(t_i-)), \qquad z'(t_i+)\in\overline{\text{co}}\, F(z(t_i+)).$$

This and (a), (b) imply that

$$\langle q(t_i), z'(t_i-)\rangle \leqq \min_{u\in U(z(t_i+))} \langle q(t_i+), f(z(t_i+), u)\rangle$$

$$\leqq \langle q(t_i+), z'(t_i+)\rangle \leqq \min_{u\in U(z(t_i))} \langle q(t_i), f(z(t_i), u)\rangle \leqq \langle q(t_i), z'(t_i-)\rangle$$

and claim (c) follows. □

Proof of Theorem 6.2. Since z is an optimal trajectory, $g'(z(1))w \geqq 0$ for all $w \in T_{R(1,C)}(z(1))$. Thus, $g'(z(1)) \in T_{R(1,C)}(z(1))^+$ and we may apply Lemma 6.8 with $p = g'(z(1))$. Set

$$q(t_{l-1}) = q(t_{l-1}+)A_{l-1} \in T_{R(t_{l-1},C)}(z(t_{l-1}))^+.$$

Then Lemma 6.8 can be applied again with $p = q(t_{l-1})$. We complete the proof using an induction argument and Lemma 6.8. □

Observe that the Lipschitz continuity of $f(x, \cdot)$ is needed to prove the local Lipschitzianity of the map $x \to f(x, U(x))$. When the control map U does not depend on x, the set-valued map $x \to f(x, U)$ is locally Lipschitzian and therefore the same proof implies Theorem 6.3.

Remark. Theorems 6.2 and 6.3 can be stated together with a higher-order condition on the adjoint vector q. However, we do not do so here to simplify the presentation of the result.

7. Examples.

Example 1. *Smooth control system.* Consider the following optimal control problem in $\mathbb{R}^2$:

$$\text{minimize } y(1)$$

over the solutions of control system

$$\begin{aligned} &x' = 1 + u(x + y^2), \qquad u \in \{0, 1\}, \\ &y' = u(2y - x), \qquad x(0) = y(0) = 0. \end{aligned} \tag{7.1}$$

Set $\bar{u} \equiv 0$. Then $z(t) = (t, 0)$ is a solution of (7.1). Moreover, $q \equiv (0, 1)$ verifies the maximum principle (5.4). On the other hand, setting $u \equiv 1$, we obtain the following Taylor expansion of the corresponding solution (x, y) of (7.1):

$$x(t) = tx'(0) + \frac{t^2}{2}x''(0) + o(t^2) = t + \frac{t^2}{2} + o(t^2),$$

$$y(t) = ty'(0) + \frac{t^2}{2}y''(0) + o(t^2) = -\frac{t^2}{2} + o(t^2).$$

Hence $z(t) + t^2(\frac{1}{2}, -\frac{1}{2}) \in R(t, 0) + o(t^2)$, and therefore $(\frac{1}{2}, -\frac{1}{2}) \in W(0, z)$. But $\langle (0, 1), (\frac{1}{2}, -\frac{1}{2})\rangle < 0$. Comparing with (5.6), we deduce that the pair $(z, \bar{u})$ is not optimal.

Example 2. *Implicit dynamical system.* Consider the following problem in $\mathbb{R}^2$:

$$\text{minimize } 2 \sin y(1) - x(1)$$

over the solutions of the implicit system

$$\dot{x}^4 + \exp(\dot{y} - 2\dot{x}) - 16x^2 - \exp(4x^2 - y^2) = 0, \quad x(0) = 0, \quad y(0) = 0. \tag{7.2}$$

Then (7.2) satisfies all the assumptions of Lemma 5.8. Observe that $z = (x, y) \equiv 0$ is a solution of (7.2). Set $q \equiv (-1, 2)$ and

$$F(0) = \{(u, v): u^4 + \exp(u - 2v) - 1 = 0\}.$$

Then for all $(u, v) \in F(0)$, $u - 2v \leqq 0$. Hence $\min\{\langle q, e\rangle : e \in F(0)\} \geqq 0$. Therefore q verifies the maximum principle (5.18)-(5.20). On the other hand, the trajectory $t \to (-t^2, -2t^2)$ is a solution of (7.2). Hence $(-1, -2) \in W(0, z)$ and $\langle (-1, 2)(+1, -2)\rangle = -3 < 0$. Consequently (5.21) does not hold and therefore the zero trajectory is not optimal.

Example 3. *Differential inclusion.* Consider the problem

$$\text{minimize } g(x(1))$$

over the solutions of the differential inclusion

$$x' \in F(x), \qquad x(0) = x_0 \tag{7.3}$$

where $F : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is a set-valued map with convex images satisfying (H_1) and $g : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function.

The high-order variations for this problem can be studied via an extension of Lie brackets to set-valued maps. Although, repeating arguments from [14], we can do it for general trajectory z of (7.3) at every point t where z is twice continuously differentiable, the calculations are quite lengthy. This is why in this example we only treat the case

$$0 \in F(x_0)$$

and the constant trajectory $z \equiv x_0$ using the ready results from [14].

From now on we assume that $0 \in F(x_0)$. To state a second-order condition for optimality we recall the following definition.

DEFINITION 7.1. Let $Q \subset F(x)$. We set

$$[F, F]_Q(x) = \{dF(x, a)b - dF(x, b)a : a, b \in Q\}.$$

The following theorem tests for optimality the constant trajectory $z \equiv x_0$.

THEOREM 7.2. *Let $A \subset dF(x_0, 0)$ be a Lipschitzian closed convex process, $Q \subset F(x_0)$ be a convex set such that*

(i) $0 \in \operatorname{r\,int} Q$;

(ii) *F is lower semicontinuously differentiable on $x_0 \times Q$ (see* [14]).

If $z \equiv x_0$ is optimal then there exists a solution q of the differential inclusion

$$-q' \in A^* q, \qquad q(1) = g'(x_0)$$

satisfying the minimum principle

$$\min_{e \in F(x_0)} \langle q(t), e\rangle = 0 \quad \text{for all } t \in [0, 1]$$

and the second-order condition

$$q(t) \in (dF(x_0, 0)Q)^+, \qquad q(t) \in ([F, F]_Q(x_0))^+$$

for all $t \in [0, 1[$.

Proof. Fix $t \in [0, 1[$. By [14, Thm. 5.2], $dF(x_0, 0)Q \subset R^\infty(t, x_0)$. From the proof of Theorem 6.1 in [14] we deduce that $[F, F]_Q(x_0) \subset R^\infty(t, x_0)$. Since $z \equiv x_0$ is optimal, $g'(x_0) \in T_{R(1, x_0)}(x_0)^+$. Theorem 4.2 ends the proof. □

Final remark. It is clear that the creation of a differential and "variational" calculus of set-valued maps (applied to reachable sets) is needed to make the field of applications broader. Special difficulties arise at all points where the trajectory tested for optimality is not continuously differentiable. Until now, this difficulty has not been overcome by any theorem in the literature concerning high-order necessary conditions. It is usually

assumed that the optimal trajectory is C^∞ (or piecewise C^∞) (see, for example, [20], [19], [4]). But, because of the Lavrentiev phenomenon, such an assumption is not reasonable. This is why we state necessary conditions here using "general" variations of reachable sets.

Acknowledgment. I thank Professor J. Zabczyk for bringing to my attention the applicability of the results to the considered impulse control problem.

REFERENCES

[1] J. P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.
[2] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.
[3] J. P. AUBIN, H. FRANKOWSKA, AND C. OLECH, *Controllability of convex processes*, SIAM J. Control Optim., 24 (1986), pp. 1192-1211.
[4] R. M. BIANCHINI AND G. STEFANI, *A high order maximum principle and controllability*, Quad. Ist. Mat. Univ. Dini, 12 (1986/87).
[5] A. BLAQUIÈRE, *Impulsive control with finite of infinite time horizon*, J. Optim. Theory Appl. (1985), pp. 431-440.
[6] S. L. CAMPBELL, *Singular Systems of Differential Equations* II, Pitman, Boston, 1982.
[7] J. H. CASE, *Economics and competitive process*, New York University Press, New York, 1979.
[8] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, New York, 1977.
[9] F. M. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[10] F. M. CLARKE AND P. LOEWEN, *State constraints in optimal control: a proximal normal analysis approach*, SIAM J. Control Optim., 25 (1987), pp. 1440-1456.
[11] A. F. FILIPPOV, *Classical solutions of differential equations with multi-valued right-hand side*, SIAM J. Control, 5 (1967), pp. 609-621.
[12] H. FRANKOWSKA, *The maximum principle for an optimal solution to a differential inclusion with end point constraints*, SIAM J. Control Optim., 25 (1987), 145-157.
[13] ———, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412-431.
[14] ———, *Local controllability of control systems with feedback*, J. Optim. Theory Appl., 2 (1989), to appear.
[15] ———, *Some inverse mapping theorems*, submitted.
[16] ———, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equation*, Appl. Math. Optim., to appear.
[17] M. R. HESTENES, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.
[18] B. KAŠKOSZ AND S. LOJASIEWICZ, *A maximum principle for generalized control systems*, *Nonlinear Analysis*, Theory, Methods Appl., 9 (1985), pp. 109-130.
[19] H. W. KNOBLOCH, *Higher Order Necessary Conditions in Optimal Control Theory*, Lecture Notes in Control and Information Sciences, A. V. Balakrishnan and M. Thoma, eds., Springer-Verlag, Berlin, New York, 1985.
[20] A. KRENER, *The high order maximal principle and its application to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256-293.
[21] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1969.
[22] G. LEITMANN, *Optimality and reachability with feedback control*, in Dynamical Systems and Microphysics, Academic Press, New York, 1982, pp. 119-141.
[23] J. POLOVINKIN, Russian doctoral disseration, Fiz. Tech. Inst. Moscow, 1986.
[24] M. REMPALA AND J. ZABCZYK, *The maximum principle for impulse control systems*, to appear.
[25] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1979.
[26] H. WAGNER, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859-903.
[27] J. WARGA, *Higher order conditions with and without Lagrange multipliers*, SIAM J. Control Optim., 24 (1986), pp. 715-730.
[28] L. C. YOUNG, *Lectures in the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.
[29] J. ZABCZYK, private communication, 1986.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 199-216, January 1989

011

NONLINEAR OBSERVER DESIGN BY OBSERVER ERROR LINEARIZATION*

XIAO-HUA XIA† AND WEI-BIN GAO†

Abstract. This paper studies the observer design problem by the observer error linearization approach for nonlinear systems with and without inputs. Necessary and sufficient conditions for the existence of the linearization transformation are derived. For nonlinear systems without inputs, the conditions are shown to be corrections to an existing result. A computation procedure and a different set of necessary and sufficient conditions based on the computation procedure are presented.

Key words. linearization, nonlinear observer, observer form, canonical forms, observability

AMS(MOS) subject classifications. 93B07, 93B10, 93B17, 93B50, 93C10, 93C35

1. Introduction.

Consider nonlinear systems of the following form:

$$\dot{x} = f(x, u), \quad x \in R^n, \quad u \in R^p, \tag{1.1a}$$

$$y = h(x), \quad y \in R^m \tag{1.1b}$$

where for each u, $f(\cdot, u) \in V(R^n)$, the set of smooth vector fields on R^n, $h(x) = (h_1(x), \cdots, h_m(x))^T$, $h_i(x) \in C^\infty(R^n)$, the set of smooth functions on R^n, $i = 1, \cdots, m$, and f is smooth with respect to u.

The observer error linearization problem is stated as follows.

Given a nonlinear system (1.1), and an initial state x^0, find (if possible) a neighborhood U of x^0, and a coordinates transformation

$$z = F(x) \quad (\text{or } x = W(z) = F^{-1}(z)) \tag{1.2}$$

defined on U, a pair of matrices (C, A) in dual Brunovsky canonical form with observability indices $k_1, \cdots, k_m$, i.e.,

$$A = \begin{array}{c} \begin{matrix} \sigma_1 & & \sigma_2 & & \cdots & & \sigma_m \end{matrix} \\ \left[\begin{array}{ccccccc} 0 & \cdots & 0 & & & & \\ 1 & & \vdots & 0 & \cdots & 0 & \\ & \ddots & \vdots & & & & \\ & 0 & 10 & & & & \\ & & 0 & \cdots & 0 & & \\ & & 1 & & & & \\ & 0 & & \ddots & \vdots & \cdots & 0 \\ & & & & 10 & & \\ \vdots & & \vdots & & & & \vdots \\ & & & & & 0 \cdots & 0 \\ & & & & & 1 & \\ 0 & & 0 & \cdots & & \ddots & \vdots \\ & & & & & & 10 \end{array}\right] \end{array} \begin{matrix} \\ \\ \sigma_1 \\ \\ \\ \\ \sigma_2, \\ \\ \\ \\ \sigma_m \end{matrix} \tag{1.3a}$$

* Received by the editors April 27, 1987; accepted for publication (in revised form) May 9, 1988. This research was supported by the Science Fund of the Chinese Academy of Science.

† The Seventh Research Division, Beijing Institute of Aeronautics and Astronautics, Beijing, China.

$$C = \begin{array}{c} \begin{matrix} \sigma_1 & & \sigma_2 & & & & \sigma_m & \end{matrix} \\ \begin{bmatrix} 0 & \cdots & 010 & \cdots & 00 & \cdots & 0 & \cdots & 00 \\ 0 & \cdots & 000 & \cdots & 01 & \cdots & 0 & \cdots & 00 \\ & \vdots & & & & \vdots & & & \\ 0 & \cdots & 000 & \cdots & 00 & \cdots & 0 & \cdots & 01 \end{bmatrix} \end{array} \tag{1.3b}$$

where $\sigma_i = \sum_{j=1}^{i} k_j$, $i = 1, \cdots, m$, and a mapping $a : h(U) \times R^p \to R^n$ such that

$$h(W(z)) = Cz, \tag{1.4}$$

$$F_* f(W(z), u) = Az + a(Cz, u) \tag{1.5}$$

for all $z \in F(U)$, $u \in R^p$, where F_* is the Jacobian of F.

If this problem is solvable, then in the new coordinates, the system (1.1) is in the form

$$\dot{z} = Az + a(y, u) = g(z, u), \tag{1.6a}$$

$$y = Cz \tag{1.6b}$$

and then an observer for system (1.1) can be constructed easily so that the observer error satisfies a linear dynamical system (for details, see [6]).

A nonlinear system in form (1.6) is said to be in observer form. This approach for nonlinear observer design was first proposed by Krener and Isidori [1] and Bestle and Zeitz [2] independently as the loose-sense mathematical dual of that proposed by Jakubczyk and Repondek [3] and Hunt and Su [4] for the exact linearization problem of nonlinear systems. Necessary and sufficient conditions in terms of differential geometry are derived in [1] for nonlinear systems without inputs when $m = 1$.

Recently, Li and Tao [5] obtained a different set of conditions described as rank conditions of matrices for single-output nonlinear time-variable systems.

For the most general cases, Krener and Respondek [6] give a very good characterization of the approach.

In this paper, we shall consider the observer error linearization problem for nonlinear systems both with and without inputs. Some new results are presented. As a consequence, a correction to Theorem 5.1 in Krener and Respondek [6] is made. Moreover, our method for dealing with systems with inputs is different from that of [6]. We will give a computation procedure for the transformation. Based on this computation procedure, a different set of necessary and sufficient conditions is obtained.

The organization of the paper is as follows. Sections 2 and 3 deal with systems without inputs. Section 2 is devoted to necessary conditions and § 3 to sufficient conditions. Based on the discussion in the previous sections, we consider systems with inputs in § 4. Section 5 contains examples and § 6 conclusions.

2. Necessary conditions. In this section and in § 3, we consider systems without inputs. Let us first analyse the problem as follows.

If the state transformation exists, then from (1.5), we have in coordinates representation

$$f(x) = W_* g(z) = \frac{\partial W}{\partial z} g(z). \tag{2.1}$$

Partially differentiating both sides of (2.1) with respect to z_{ij} results in

$$\frac{\partial f(x)}{\partial z_{ij}} = \frac{\partial}{\partial z_{ij}}\left(\frac{\partial W}{\partial z}\right) g(z) + \frac{\partial W}{\partial z} \cdot \frac{\partial g(z)}{\partial z_{ij}}, \tag{2.2}$$

but

$$\frac{\partial f(x)}{\partial z_{ij}} = \frac{\partial f}{\partial x} \cdot \frac{\partial W}{\partial z_{ij}}, \tag{2.3}$$

$$\frac{\partial}{\partial z_{ij}}\left(\frac{\partial W}{\partial z}\right) g(z) = \frac{\partial}{\partial x}\left(\frac{\partial W}{\partial z_{ij}}\right)\frac{\partial W}{\partial z} g(z) = \frac{\partial}{\partial x}\left(\frac{\partial W}{\partial z_{ij}}\right) f(x), \tag{2.4}$$

$$\frac{\partial W}{\partial z} \cdot \frac{\partial g(z)}{\partial z_{ij}} = \begin{cases} \dfrac{\partial W}{\partial z_{ij+1}}, & i = 1, \cdots, m; \quad j = 1, \cdots, k_i - 1, \\ \dfrac{\partial W}{\partial z} \cdot \dfrac{\partial a}{\partial z_{ik_i}}, & i = 1, \cdots, m; \quad j = k_i. \end{cases} \tag{2.5}$$

By substituting (2.3)-(2.5) for (2.2), we obtain

$$\frac{\partial W}{\partial z_{ij+1}} = \frac{\partial f}{\partial x} \cdot \frac{\partial W}{\partial z_{ij}} - \frac{\partial}{\partial x}\left(\frac{\partial W}{\partial z_{ij}}\right) f, \quad i = 1, \cdots, m; \quad j = 1, \cdots, k_i - 1, \tag{2.6}$$

$$\frac{\partial W}{\partial z} \cdot \frac{\partial a}{\partial z_{ik_i}} = \frac{\partial f}{\partial x} \cdot \frac{\partial W}{\partial z_{ik_i}} - \frac{\partial}{\partial x}\left(\frac{\partial W}{\partial z_{ik_i}}\right) f, \qquad i = 1, \cdots, m; \tag{2.7}$$

in other words,

$$\frac{\partial W}{\partial z_{ij+1}} = \mathrm{ad}_{(-f)} \frac{\partial W}{\partial z_{ij}}, \quad i = 1, \cdots, m; \quad j = 1, \cdots, k_i - 1, \tag{2.8}$$

$$\frac{\partial W}{\partial z} \cdot \frac{\partial a}{\partial z_{ik_i}} = \mathrm{ad}_{(-f)} \frac{\partial W}{\partial z_{ik_i}}, \qquad i = 1, \cdots, m. \tag{2.9}$$

These can be rewritten as

$$\frac{\partial W}{\partial z_{ij+1}} = \mathrm{ad}^j_{(-f)} \frac{\partial W}{\partial z_{i1}}, \quad i = 1, \cdots, m; \quad j = 1, \cdots, k_i - 1, \tag{2.10}$$

$$\frac{\partial W}{\partial z} \cdot \frac{\partial a}{\partial z_{ik_i}} = \mathrm{ad}^{k_i}_{(-f)} \frac{\partial W}{\partial z_{i1}}, \qquad i = 1, \cdots, m. \tag{2.11}$$

On the other hand, from (1.4) we have

$$h_p(W(z)) = c_p z, \qquad p = 1, \cdots, m \tag{2.12}$$

where $c_p = (0 \cdots 010 \cdots 0)$ with the unit in the $(\sum_{i=1}^{p} k_i)$th position.

Partially differentiating both sides of (2.12) with respect to z_{ij}, we may have

$$\langle dh_p, \partial W / \partial z_{ij} \rangle = \delta_{i,p} \cdot \delta_{j,k_i}, \quad i = 1, \cdots, m; \quad p = 1, \cdots, m; \quad j = 1, \cdots, k_i \tag{2.13}$$

where δ is the Kronecker delta. By (2.8) and the Leibniz rule, we see that

$$\begin{aligned}\langle dh_p, \partial W/\partial z_{ij}\rangle &= \langle dh_p, \mathrm{ad}_{(-f)}\partial W/\partial z_{ij-1}\rangle \\ &= L_{(-f)}\langle dh_p, \partial W/\partial z_{ij-1}\rangle - \langle L_{(-f)}(dh_p), \partial W/\partial z_{ij-1}\rangle \\ &= \langle L_f(dh_p), \partial W/\partial z_{ij-1}\rangle, \quad i, p = 1, \cdots, m; \quad j = 2, \cdots, k_i,\end{aligned} \tag{2.14}$$

so formulae (2.13) imply, for $i = 1, \cdots, m$,

$$\langle L_f^{j-1}(dh_p), \partial W/\partial z_{i1}\rangle = \delta_{i,p} \cdot \delta_{j,k_i}, \quad p = 1, \cdots, m; \quad j = 1, \cdots, k_i. \tag{2.15}$$

Thus, the problem is reduced to the investigation of the solvability of (2.10), (2.11), and (2.15). We may formulate the analysis above as a theorem.

THEOREM 2.1. *The state transformation* (1.2) *transforms the system* (1.1a)-(1.1b) *into system* (1.6a)-(1.6b) *if and only if W and a satisfy* (2.10), (2.11), *and* (2.15).

Remark 2.2. Note that (2.15) are linear algebraic equations in $\partial W/\partial z_{i1}$. For each i, there are mk_i equations altogether. If $mk_i > n$, then these equations are overdetermined, and if $mk_i < n$, they are underdetermined. Therefore, $\partial W/\partial z_{i1}$ may not be uniquely obtained. When the observability indices are identical, i.e., $k_1 = k_2 = \cdots = k_m$, then $\partial W/\partial z_{i1}$ is uniquely determined.

Yet, the solvability of these linear algebraic equations is always guaranteed by a simple consequence of the necessary conditions developed below.

THEOREM 2.3. *The observer error linearization problem is solvable only if there exist m-tuple of integers* $(k_1, \cdots, k_m)$, $k_1 \geqq k_2 \geqq \cdots \geqq k_m > 0$, *and* $\sum_{i=1}^m k_i = n$, *such that we have the following*:

(i) *If we denote* (*with a possible reordering of the* h_i*'s*)

$$Q = \{L_f^{j-1}(dh_i): i = 1, \cdots, m; j = 1, \cdots, k_i\}, \tag{2.16}$$

then

$$\dim \operatorname{span} Q = n \tag{2.17}$$

in a neighborhood of x^0.

(ii) *If we denote*

$$Q_j = \{L_f^{k-1}(dh_i): i = 1, \cdots, m; k = 1, \cdots, k_j\} - \{L_f^{k_j-1}(dh_j)\} \tag{2.18}$$

for $j = 1, \cdots, m$, *then*

$$\dim \operatorname{span} Q_j = \dim \operatorname{span} Q \cap Q_j \tag{2.19}$$

for $j = 1, \cdots, m$.

Remark 2.4. Condition (i) means that the nonlinear system is observable in some sense.

Remark 2.5. Condition (ii) means that in representing those 1 forms in $Q_i - Q \cap Q_i$ as linear combinations of the 1 forms in Q, the coefficients attached to the 1 forms in $Q - Q \cap Q_i$ are zeros. This is a dual result of 1 in Hunt and Su [4].

Proof of Theorem 2.3. If $z = F(x)$ or $x = W(z)$ is the state transformation, then

$$\frac{\partial W}{\partial z} = \left(\frac{\partial W}{\partial z_{11}} \cdots \frac{\partial W}{\partial z_{1k_1}} \cdots \frac{\partial W}{\partial z_{m1}} \cdots \frac{\partial W}{\partial z_{mk_m}}\right)$$

is nonsingular. Denote the matrix $O(x)$ as

$$O(x)=\begin{bmatrix} dh_1 \\ L_f(dh_1) \\ \vdots \\ L_f^{k_1-1}(dh_1) \\ \vdots \\ dh_m \\ L_f(dh_m) \\ \vdots \\ L_f^{k_m-1}(dh_m) \end{bmatrix} \tag{2.20}$$

and consider the matrix product $O(x)\cdot \partial W/\partial z$.

From (2.10) and the Leibniz rule, it is easy to see that

$$O(x)\cdot\frac{\partial W}{\partial z}=\begin{bmatrix} 0 & 1 & & & & & & \\ & \ddots & x & 0 & \cdots & & 0 & \\ 1 & x & x & & & & & \\ & & & 0 & 1 & & & \\ & 0 & & \ddots & x & \cdots & 0 & \\ & & & 1 & x & x & & \\ & \vdots & & & \vdots & & & \\ & & & & & & 0 & 1 \\ & 0 & & 0 & & & \ddots & x \\ & & & & & & 1 & x & x \end{bmatrix}. \tag{2.21}$$

The matrix in the right-hand side of (2.21) is nonsingular. This implies (i).

Now we prove (ii). Obviously,

$$\text{span } Q_i \supseteq \text{span } Q \cap Q_i, \qquad i=1,\cdots,m, \tag{2.22}$$

so, if we can prove that $\dim \text{span } Q_i \leqq \dim \text{span } Q\cap Q_i$, we then have (ii).

On the one hand, because of (i), it is clear that

$$\dim \text{span } Q\cap Q_i = ik_i + k_{i+1}+\cdots+k_m-1 \tag{2.23}$$

for $i=1,\cdots,m$. On the other hand, by (2.13) and repeated use of the Leibniz rule, we may have

$$\langle L_f^{k-1}(dh_p)\partial W/\partial z_{js}\rangle = 0 \tag{2.24}$$

for $i=1,\cdots,m$; $k=1,\cdots,k_i-1$; $j=1,\cdots,i$; $s=1,\cdots,k_j-k_{i+1}$. That is, $\partial W/\partial z_{11},\cdots,\partial W/\partial z_{1k_1-k_i+1},\cdots,\partial W/\partial z_{i-11},\cdots,\partial W/\partial z_{i-1k_{i-1}-k_i+1}$, $\partial W/\partial z_{i1}$ annihilate the one forms in Q_i. Because of the independence of $\partial W/\partial z_{ij}$'s we deduce that

$$\begin{aligned} \dim \text{span } Q_i &\leqq n-(k_1-k_i)-\cdots-(k_{i-1}-k_i)-1 \\ &= ik_i+k_{i+1}+\cdots+k_m-1. \end{aligned} \tag{2.25}$$

So, from (2.23)

$$\dim \text{span } Q_i \leqq \dim \text{span } Q\cap Q_i.$$

This completes the proof. □

COROLLARY 2.6. *If conditions* (i) *and* (ii) *of Theorem* 2.3 *hold, then the linear algebraic equations* (2.15) *are solvable.*

Proof. By (ii) and Remark 2.5, we know that (2.15) are in fact equivalent to the following equations (write g for $\partial W/\partial z_{i1}$):

$$\begin{aligned}
&\langle dh_1, g\rangle = 0 \cdots \langle dh_i, g\rangle = 0, && \langle dh_{i+1}, g\rangle = 0 \cdots \langle dh_m, g\rangle = 0\\
&\langle L_f dh_1, g\rangle = 0 \cdots \langle L_f dh_i, g\rangle = 0, && \langle L_f dh_{i+1}, g\rangle = 0 \cdots \langle L_f dh_m, g\rangle = 0\\
&\quad\vdots \qquad\qquad \vdots && \quad\vdots \qquad\qquad \vdots\\
&\langle L_f^{k_i-1} dh_1, g\rangle = 0 \cdots \langle L_f^{k_i-1} dh_i, g\rangle = 1 && \langle L_f^{k_{i+1}} dh_{i+1}, g\rangle = 0 \cdots \langle L_f^{k_m-1} dh_m, g\rangle = 0.
\end{aligned} \tag{2.26}$$

Solutions always exist for these $ik_i + k_{i+1} + \cdots + k_m$ equations, since $\dim \operatorname{span}[Q_i \cup \{L_f^{k_i-1}(dh_i)\}] = ik_i + k_{i+1} + \cdots + k_m$. □

Now let $g^1, \cdots, g^m$ be solutions to (2.15), and define a matrix as follows:

$$\tilde{Q}(x) = (g^1 \operatorname{ad}_{(-f)} g^1 \cdots \operatorname{ad}_{(-f)}^{k_1-1} \cdots g^m \operatorname{ad}_{(-f)} g^m \cdots \operatorname{ad}_{(-f)}^{k_m-1} g^m). \tag{2.27}$$

As in the argument of Theorem 2.3, we have that $\tilde{Q}(x)$ is nonsingular if the system is observable (that is, if $O(x)$ is nonsingular). Thus from (2.11) we have

$$\frac{\partial a}{\partial z_{ik_i}} = \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_i} g^i. \tag{2.28}$$

Another necessary condition can be derived by (2.28).

THEOREM 2.7. *A necessary condition for the system* (1.1) *to admit an observer form* (1.6) *is that there exist solutions* $g^1, \cdots, g^m$ *of* (2.15) *such that*

$$\operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_1} g^1 \\ \vdots \\ \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_m} g^m \\ h(x) \end{bmatrix} = m. \tag{2.29}$$

Proof. If the state transformation $x = W(z)$ exists, then we have

$$y = (z_{1k_1} \cdots z_{mk_m})^T = h(x). \tag{2.30}$$

However, from the previous remark, (2.10), and (2.11), we have

$$\frac{\partial a}{\partial z_{ik_i}} = \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_i} g^i = b_i(y) = b_i(h(x)) \tag{2.31}$$

where

$$\begin{aligned}
b_i(y) &= (b_{00}^i(y) \cdots b_{mk_m-1}^i(y))^T \\
&= \frac{\partial a(y)}{\partial y_i} = \frac{\partial a(z_{1k_1}, \cdots, z_{mk_m})}{\partial z_{ik_i}}
\end{aligned} \tag{2.32}$$

is dependent only on y, so

$$\operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_1} g^1 \\ \vdots \\ \tilde{Q}^{-1}(x) \operatorname{ad}_{(-f)}^{k_m} g^m \\ h(x) \end{bmatrix} = \operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} b_1(h(x)) \\ \vdots \\ b_m(h(x)) \\ h(x) \end{bmatrix} = \operatorname{rank} \frac{\partial h}{\partial x} = m. \qquad \square$$

3. Sufficient conditions. We have already seen from the previous section that the observer error linearization problem is transformed to the study of the integrability of the partial differential equations (2.10) and (2.11) and the solvability of the linear algebraic equations (2.15). By considering (2.15) and (2.11), we have obtained the

necessary conditions. In this section we turn our attention to the integrability of the partial differential equations (2.10) and (2.11). As remarked by Krener and Respondek [6], the integrability of (2.10) implies the integrability of (2.11). We will show that the converse is also true. And the integrability of (2.10) and of (2.11) leads to sufficient conditions for the observer error linearization problem.

First, consider (2.10); we then have Theorem 3.1.

THEOREM 3.1. *The observer error linearization problem is solvable if and only if there exist m-tuple of integers* $(k_1, \cdots, k_m)$, $k_1 \geqq k_2 \geqq \cdots \geqq k_m > 0$, *and* $\sum_{i=1}^{m} k_i = n$, *such that we have the following*:

(i) *Conditions* (i) *and* (ii) *in Theorem* 2.3 *hold.*

(ii) *There exists a mapping* ϕ *of some open set* V *of* R^n *onto a neighborhood* U *of* x^0 *and vector fields* $g^1, \cdots, g^m$ *satisfying*

$$L_{g_i} L_f^{l-1}(h_j) = \delta_{i,j} \cdot \delta_{l,k_j}, \quad i = 1, \cdots, m, \quad l = 1, \cdots, k_i, \quad j = 1, \cdots, m \tag{3.1}$$

such that

$$\frac{\partial \phi}{\partial z} = (g^1 \operatorname{ad}_{(-f)} g^1 \cdots \operatorname{ad}_{(-f)}^{k_1 - 1} g^1 \cdots g^m \operatorname{ad}_{(-f)} g^m \cdots \operatorname{ad}_{(-f)}^{k_m - 1} g^m) \circ \phi(z) \tag{3.2}$$

for all $z \in V$.

Remark 3.2. The g^i's satisfying (3.1) can always be found if Theorem 3.1(i) holds. As a matter of fact, we need to solve only the set of equations

$$\langle L_f^{l-1}(dh_j), g^i \rangle = \delta_{i,j} \cdot \delta_{l,k_j}, \quad i, j = 1, \cdots, m; \quad l = 1, \cdots, k_i. \tag{3.3}$$

These are just the linear algebraic equations (2.15). Thus, Corollary 2.6 and (1) imply the solvability of (3.3). □

Remark 3.3. Equations (3.2) are, in fact, (2.10). So, the theorem implies that the integrability of (2.11) is guaranteed by the integrability of (2.10). Also, we can show the converse (see Xia and Gao [7]).

Proof of Theorem 3.1. The necessity follows easily from Remarks 3.2 and 3.3 and Theorem 2.3.

Sufficiency. Suppose (1), (2) hold. As in the proof of Theorem 2.3, we may immediately note that the matrix

$$\begin{bmatrix} dh_1 \\ L_f(dh_1) \\ \vdots \\ L_f^{k_{m^1}}(dh_1) \\ \vdots \\ dh_m \\ L_f(dh_m) \\ \vdots \\ L_f^{k_m - 1}(dh_m) \end{bmatrix} \cdot [g^1 \operatorname{ad}_{(-f)} g^1 \cdots \operatorname{ad}_{(-f)}^{k_1 - 1} g^1 \cdots g^m \operatorname{ad}_{(-f)} g^m \cdots \operatorname{ad}_{(-f)}^{k_m - 1} g^m]$$

has rank n in a neighborhood U of x^0. Therefore, the vector fields $\operatorname{ad}_{(-f)}^j g^i$, $i = 1, \cdots, m$; $j = 0, \cdots, k_i - 1$ are linearly independent in the neighborhood U of x^0.

Let $z^0 \in R^n$, such that $\phi(z^0) = x^0$. From the linear independence of the vector fields on the right-hand side of (3.2), we deduce that ϕ has rank n at z^0, i.e., that ϕ is a diffeomorphism of a neighborhood of z^0 onto a neighborhood of x^0.

Set $W=\phi$, or $F=\phi^{-1}$, and

$$F_* f\circ F^{-1}(z)=\hat{f}_{11}\frac{\partial}{\partial z_{11}}+\cdots+\hat{f}_{1k_1}\frac{\partial}{\partial z_{1k_1}}+\cdots+\hat{f}_{m1}\frac{\partial}{\partial z_{m1}}+\cdots+\hat{f}_{mk_m}\frac{\partial}{\partial z_{mk_m}}. \tag{3.4}$$

By (2), the mapping ϕ is such that

$$\phi_*\left(\frac{\partial}{\partial z_{ij+1}}\right)\circ\phi^{-1}(x)=\mathrm{ad}^j_{(-f)}g^i(x)$$

so that

$$F_*\,\mathrm{ad}^j_{(-f)}g^i\circ F^{-1}(z)=\frac{\partial}{\partial z_{ij+1}} \tag{3.5}$$

for $i=1,\cdots,m$; $j=0,\cdots,k_i-1$.

Using (3.4) and (3.5), we obtain, for $i=1,\cdots,m$; $j=0,\cdots,k_i-2$,

$$\begin{aligned}\frac{\partial}{\partial z_{ij+2}}&=F_*\,\mathrm{ad}^{j+1}_{(-f)}g^i\circ F^{-1}(z)=F_*[-f,\mathrm{ad}^j_{(-f)}g^i]\circ F^{-1}(z)\\&=[-F_*f\circ F^{-1}(z),F_*\,\mathrm{ad}^j_{(-f)}g^i\circ F^{-1}(z)]\\&=-\left[\hat{f}_{11}\frac{\partial}{\partial z_{11}}+\cdots+\hat{f}_{mk_m}\frac{\partial}{\partial z_{mk_m}},\frac{\partial}{\partial z_{ij+1}}\right]\\&=\frac{\partial\hat{f}_{11}}{\partial z_{ij+1}}\frac{\partial}{\partial z_{11}}+\cdots+\frac{\partial\hat{f}_{mk_m}}{\partial z_{ij+1}}\frac{\partial}{\partial z_{mk_m}}.\end{aligned}$$

Because of the linear independence of the $\partial/\partial z_{ij}$'s, this implies

$$\frac{\partial f_{lk}}{\partial z_{ij+1}}=\delta_{i,l}\cdot\delta_{k,j+2}. \tag{3.6}$$

We then deduce that $\hat{f}_{11},\cdots,\hat{f}_{m1}$ depend only on $z_{1k_1},\cdots,z_{mk_m}$, and that $\hat{f}_{ij}$ for $i=1,\cdots,m$; $j=2,\cdots,k_i$, is such that $\hat{f}_{ij}-z_{ij-1}$ depend only on $z_{1k_1},\cdots,z_{mk_m}$. In other words, we have

$$F_*f\circ F^{-1}(z)=Az+a(z_{1k_1},\cdots,z_{mk_m})$$

where a is a suitable mapping of z_{ik_i}, and this shows that condition (1.5) holds.

Moreover, since, by (3.1).

$$L_{\mathrm{ad}^{l-1}_{(-f)}g^i}h_j=\delta_{i,j}\cdot\delta_{l,k_i}$$

for $i,j=1,\cdots,m$; $l=1,\cdots,k_i$, we have that

$$\frac{\partial h_j\circ F^{-1}}{\partial z_{il}}=\delta_{i,j}\cdot\delta_{j,k_i}$$

for $i,j=1,\cdots,m$; $l=1,\cdots,k_i$. In other words,

$$\frac{\partial}{\partial z}(h\circ F^{-1}(z))=C.$$

This implies that condition (1.4) holds. □

The integrability of the partial differential equations (3.2) may be expressed in terms of a property of the vector fields on the right-hand side (see Spivak [8]).

THEOREM 3.4. *The observer error linearization problem is solvable if and only if* (i) *and* (ii) *in Theorem* 2.3 *hold and there exist vector fields* $g^1,\cdots,g^m$ *satisfying* (3.1)

such that

$$[\mathrm{ad}^{k}_{(-f)}g^{i}, \mathrm{ad}^{l}_{(-f)}g^{j}]=0 \tag{3.7}$$

for $i, j=1, \cdots, m$; $k=0, \cdots, k_i-1$; $l=0, \cdots, k_j-1$.

Now, several consequences of the results above are immediate.

COROLLARY 3.5. *A nonlinear system* (1.1a)-(1.1b) *admits an observer form* (1.6a)-(1.6b) *with the observability indices in the pair* (C, A) *all identical if and only if*:

(i) dim span $Q=n$;

(ii) $[\mathrm{ad}^{k}_{(-f)}g^{i}, \mathrm{ad}^{l}_{(-f)}g^{j}]=0$,

for $i, j=1, \cdots, m$; $k, l=0, \cdots, k_i-1$, *where* g^i *is the vector field determined by* (3.1).

Proof. If we notice that when $k_1=k_2=\cdots=k_m$, $Q \cap Q_i=Q_i$, then the proof is trivial. □

Also, if we prolong the linear algebraic equations such that g^i can be uniquely defined, a sufficient condition follows.

COROLLARY 3.6. *Let* g^j *be the vector field defined by the following n equations*:

$$L_{g^j}L_f^{l-1}h_i=\delta_{i,j}\cdot\delta_{l,k_i}, \quad i=1, \cdots, m; \quad l=1, \cdots, k_i \tag{3.8}$$

for $j=1, \cdots, m$, *then the observer error linearization problem is solvable if conditions* (3.7) *hold.*

Remark 3.7. Now we can see that the conditions in Theorem 5.1 of [6] are sufficient but not necessary, as a counterexample in the next section will show.

Remark 3.8. The partial differential equations satisfying condition (3.7) can be integrated by standard methods as outlined in [8]. The computation algorithm in Hunt and Su [4] can also be used to solve these equations.

The state transformation can also be obtained by considering the integrability of (3.11). This has been done in Xia and Gao [7] for nonlinear systems without inputs. In [7], a computation procedure is proposed, and based on the computation procedure, we obtained a different set of necessary and sufficient conditions. For comparison with the previous result and for the development of the next section, we briefly review some results in [7].

A COMPUTATION PROCEDURE.

(i) Compute $O(x)$ defined in (2.20).

(ii) Choose solutions $g^1, \cdots, g^m$ of (2.26) or (2.15), and compute $\tilde{Q}(x)$ defined in (2.27), and b_i's defined in (2.31).

(iii) If b_i's are functions of y (i.e., $z_{1k_1}, \cdots, z_{mk_m}$), solve the following equations:

$$\frac{\partial a}{\partial y_i}=b_i(y). \tag{3.9}$$

(iv) Compute the state transformation

$$z=F(x)=\begin{bmatrix} z_{11}(x) \\ \vdots \\ z_{1k_1}(x) \\ \vdots \\ z_{m1}(x) \\ \vdots \\ z_{mk_m}(x) \end{bmatrix} \tag{3.10}$$

as follows:

$$z_{ik_i}(x)=h_i(x) \quad \text{for } i=1, \cdots, m, \tag{3.11a}$$

$$z_{ij}(x) = L_f z_{ij+1}(x) - a_{ij}(h(x)) \tag{3.11b}$$

for $i = 1, \cdots, m$; $j = k_i - 1, k_i - 2, \cdots, 2, 1$.

(v) Redefine $a_{i0}(x_{1k_1}, \cdots, x_{mk_m}) = a_{i0}(y) = L_f z_{i1}(x)$, for $i = 1, \cdots, m$.

Before we introduce the results, some conventions are necessary.

For multivalued functions $g = (g_1, \cdots, g_m)$, $g_i = g_i(y_1, \cdots, y_m) \in R^n$, we denote the following $mn \times m$ matrix as $[\partial g / \partial y]$, i.e.,

$$\left[\frac{\partial g}{\partial y}\right] = \begin{bmatrix} \dfrac{\partial g_1}{\partial y_1} & \dfrac{\partial g_1}{\partial y_2} & \cdots & \dfrac{\partial g_1}{\partial y_m} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial g_m}{\partial y_1} & \dfrac{\partial g_m}{\partial y_2} & \cdots & \dfrac{\partial g_m}{\partial y_m} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{matrix} n \\ \vdots \\ n \end{matrix}. \tag{3.12}$$

An $mn \times m$ matrix A partitioned as follows will be denoted as $\bar{A}$,

$$\bar{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1m} \\ A_{21} & A_{22} & \cdots & A_{2m} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mm} \\ 1 & 1 & & 1 \end{bmatrix} \begin{matrix} n \\ n \\ \vdots \\ n \\ mn \times m \end{matrix} = (a_{ij})_{mn \times m} = A. \tag{3.13}$$

An $mn \times m$ matrix A is called block symmetric if A_{ij}'s in the partitioned $\bar{A}$ in (3.13) satisfy

$$A_{ij} = A_{ji}, \qquad i, j = 1, \cdots, m$$

and the block symmetrical property of an $mn \times m$ matrix A is denoted as

$$\bar{A} = \bar{A}^T.$$

Now, we are ready to introduce our results.

THEOREM 3.9. *The system* (1.1) *admits an observer form* (1.6) *if and only if there exist m-tuple of integers* $(k_1, \cdots, k_m)$, $k_1 \geqq k_2 \geqq \cdots \geqq k_m > 0$, *and* $\sum_{i=1}^m k_i = n$, *such that*:

(1) *Conditions* (i) *and* (ii) *in Theorem* 2.3 *hold.*

(2) *There exist solutions* $g^1, \cdots, g^m$ *to* (2.26) *such that condition* (2.29) *in Theorem* 2.7 *holds.*

(3) *If we denote an* $mn \times m$ *matrix* $S(x)$ *as*

$$S(x) = \frac{\partial}{\partial x}\begin{bmatrix} \tilde{Q}^{-1}(x)\, \mathrm{ad}_{(-f)}^{k_1} g^1 \\ \vdots \\ \tilde{Q}^{-1}(x)\, \mathrm{ad}_{(-f)}^{k_m} g^m \end{bmatrix} \left(\frac{\partial h}{\partial x}\right)^T \left[\frac{\partial h}{\partial x}\left(\frac{\partial h}{\partial x}\right)^T\right]^{-1}, \tag{3.14}$$

then $S(x)$ *is block symmetric, i.e.,*

$$\overline{S(x)} = (\overline{S(x)})^T. \tag{3.15}$$

Remark 3.10. This theorem guarantees the computability of the procedure. We see that step (iii) is the most difficult part of the procedure. Under the conditions of this theorem, the partial differential equations (3.9) are usually referred to as exact equations. Solutions of exact equations are easier to obtain. So, from the computational point of view, the method here is more convenient than the method we used when solving the partial differential equations (3.2).

Remark 3.11. From this theorem, it is easily seen that the integrability of (2.11) implies the integrability of (2.10).

Remark 3.12. The proof of this theorem can be found in [7]. From the proof, we can see that there are some abundances in (3.9). In fact, equations concerning a_{i0}, $i=1,\cdots,m$, in (3.9) can be dropped. Thus, we need to check appropriate conditions of lower-dimensional matrices. For details see [7] and the next section.

4. Systems with inputs. In this section, we consider systems with inputs. From the previous discussions, it is easily known that (2.10), (2.11), and (2.16) hold in this case, except that L_f and $\mathrm{ad}_{(-f)}$ are replaced by $L_{f(\cdot,u)}$ and $\mathrm{ad}_{(-f(\cdot,u))}$, respectively. Throughout this section, for notational simplicity, we will assume that L_f and $\mathrm{ad}_{(-f)}$ stand for $L_{f(\cdot,u)}$ and $\mathrm{ad}_{(-f(\cdot,u))}$, respectively. Now the matrices O defined in (2.20) and $\tilde{Q}$ defined in (2.27), respectively, are u dependent. A u-dependent proposition is true if it is true for any u. Thus, the necessary conditions developed in § 2 are also valid here. We remark that the nonsingularity of the matrix $O(x,u)$ relates to a kind of observability property of the system (see [9]). To derive sufficient conditions, we note that the mapping a must satisfy some extra requirement because it is now u dependent. In this section, we will give a modification of the computation procedure introduced above.

First, we check each step in the computation procedure. Steps (i) and (ii) will present no problem. Step (iii) needs some modification. This is because there will be certain restrictions on the dependence of the mapping $a(y,u)$ on the inputs u to perform step (iv).

We explicitly do as follows. From step (iv), we have

$$\begin{aligned} z_{ik_i-j} &= L_f^j h_i(x) - L_f^{j-1} a_{ik_i-1} - \cdots - L_f a_{ik_i-j+1} - a_{ik_i-j} \\ &= L_f^j h_i(x) - L_f^{j-2} \frac{\partial a_{ik_i-1}}{\partial y} \frac{\partial h}{\partial x} \cdot f - \cdots \frac{\partial a_{ik_i-j+1}}{\partial y} \frac{\partial h}{\partial x} \cdot f - a_{ik_i-j} \end{aligned}$$

for $i=1,\cdots,m$; $j=0,1,\cdots,k_i-1$. Now, since

$$\frac{\partial a_{ij}}{\partial y} = \left(\frac{\partial a_{ij}}{\partial y_1} \cdots \frac{\partial a_{ij}}{\partial y_m}\right)^T = (b_{ij}^1 \cdots b_{ij}^m)^T$$

we have

$$\begin{aligned} z_{ik_i-j} &= L_f^j h_i(x) - \sum_{s=1}^{j-1} L_f^{j-s-1} \left((b_{ik_i-s}^1 \cdots b_{ik_i-s}^m)^T \frac{\partial h}{\partial x} f \right) - a_{ik_i-j} \\ &\triangleq \xi_{ik_i-j} - a_{ik_i-j}. \end{aligned}$$

If we denote

$$b_{p+i}(x,u) = \frac{\partial}{\partial u_i} (\xi_{10} \cdots \xi_{1k_1-1} \cdots \xi_{m0} \cdots \xi_{mk_m-1})^T \tag{4.1}$$

for $i=1,\cdots,m$, then we will have the extra equations the mapping a must satisfy

$$\frac{\partial a}{\partial u_i} = b_{p+i}(x,u) \tag{4.2}$$

for $i=1,\cdots,m$. Therefore, we have a modified computation procedure.

(i) Compute $O(x,u)$ defined in (2.20).

(ii) Choose solutions $g^1,\cdots,g^m$ of (2.26) or (2.15), and compute $\tilde{Q}(x,u)$ defined in (2.27) and b_i's defined in (2.31) and (4.1).

(iii) If b_i's are functions of y and u, then solve the following equations:

$$\frac{\partial a}{\partial y_i} = b_i(y, u), \qquad i = 1, \cdots, p, \tag{4.3a}$$

$$\frac{\partial a}{\partial u_j} = b_{p+j}(y, u), \qquad j = 1, \cdots, m. \tag{4.3b}$$

(iv) Compute the state transformation as follows:

$$z_{ik_i}(x) = h_i(x) \quad \text{for } i = 1, \cdots, m, \tag{4.4a}$$

$$z_{ij}(x) = L_f z_{ij+1}(x) - a_{ij}(h(x), u) \tag{4.4b}$$

for $i = 1, \cdots, m$; $j = k_i - 1, \cdots, 1$.

(v) Redefine $a_{i0}(y, u) = L_f z_{i1}(x) \triangleq \lambda_i(x, u)$, for $i = 1, \cdots, m$.

Now, we have to show that if a transformation exists (it is easily seen that all the above steps are computable), then the computation procedure shown above does give the transformation. Moreover, for the computation of step (iii), we must find conditions that ensure the integrability of (4.3), and we have to show that the z_{ij}'s computed in step (iv) are independent of the inputs u. This leads to the sufficient conditions below.

THEOREM 4.1. *The system* (1.1) *admits an observer form* (1.6) *if and only if there exist m-tuple of integers* $(k_1, \cdots, k_m)$, $k_1 \geqq k_2 \geqq \cdots \geqq k_m > 0$, and $\sum_{i=1}^{m} k_i = n$, *such that*:

(1) *Conditions* (i) *and* (ii) *in Theorem* 2.3 *hold.*

(2) *There exist solutions* $g^1, \cdots, g^m$ *to* (2.26) *such that*

$$\operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \lambda_1(x, u) \\ \vdots \\ \lambda_m(x, u) \\ b_1(x, u) \\ \vdots \\ b_{m+p}(x, u) \\ h(x) \end{bmatrix} = m \tag{4.5}$$

where the b_i*'s are from* (2.31) *and* (4.1), *and the* λ_i*'s are from step* (v).

(3) *If we denote an* $(m+p)n \times (m+p)$ *matrix* $S(x, u)$ *as*

$$S(x, u) = \left[\frac{\partial}{\partial x} \begin{bmatrix} b_1(x, u) \\ \vdots \\ b_{m+p}(x, u) \end{bmatrix} \left(\frac{\partial h}{\partial x} \right)^T \left[\frac{\partial h}{\partial x} \left(\frac{\partial h}{\partial x} \right)^T \right]^{-1} \middle| \frac{\partial}{\partial u} \begin{bmatrix} b_1(x, u) \\ \vdots \\ b_{m+p}(x, u) \end{bmatrix} \right], \tag{4.6}$$

then $S(x, u)$ *is block symmetric, i.e.,*

$$\overline{S(x, u)} = (\overline{S(x, u)})^T. \tag{4.7}$$

Proof. Necessity. We only prove (3); the other conditions are similarly proved, as are the necessary conditions proved in § 2.

If the state transformation exists, then from the above analysis, the following equations are satisfied by the differentiable function $a(x, u)$,

$$\frac{\partial a}{\partial y_i} = b_i(y, u) \quad \text{for } i = 1, \cdots, m, \tag{4.8a}$$

$$\frac{\partial a}{\partial u_j} = b_{m+j}(y, u) \quad \text{for } j = 1, \cdots, p \tag{4.8b}$$

where, by abuse of notation,

$$b_i(y, u) = b_i(h(x), u) = b_i(x, u)$$

and the last terms of the equality above are computed in step (ii) of the procedure. By condition (2), they are functions of y and u.

So, by Lemma 4.1 of Xia and Gao [7],

$$\overline{\left[\frac{\partial b}{\partial y}\ \frac{\partial b}{\partial u}\right]} = \left(\left[\overline{\frac{\partial b}{\partial y}\ \frac{\partial b}{\partial u}}\right]\right)^T. \tag{4.9}$$

But, again by abuse of notation,

$$\frac{\partial b_i}{\partial y}\frac{\partial h}{\partial x} = \frac{\partial b_i}{\partial x}.$$

Multiplying both sides by $(\partial h/\partial x)^T$, and noting that $\partial h/\partial x(\partial h/\partial x)^T$ is an $m \times m$ nonsingular matrix, we have

$$\frac{\partial b_i}{\partial y} = \frac{\partial b_i}{\partial x}\left(\frac{\partial h}{\partial x}\right)^T\left[\frac{\partial h}{\partial x}\left(\frac{\partial h}{\partial x}\right)^T\right]^{-1}. \tag{4.10}$$

From (4.10), we see immediately that

$$\left[\frac{\partial b}{\partial y}\ \frac{\partial b}{\partial u}\right] = S(x, u). \tag{4.11}$$

Hence, (4.9) gives (4.7). This completes the proof of necessity.

Sufficiency. By (2) we know that the b_i's are functionally dependent with $h(x)$. This implies that the b_i's are functions of $h(x)$, or y. And by (3), the same argument as in the proof of necessity will show that (4.9) holds. This in turn implies the integrability of (4.3). Moreover, a direct computation shows that the z_{ij}'s are independent of the inputs u.

Thus, we need only to show that the transformation $z = F(x)$ given by the above computation procedure does transform system (1.5) into (1.1).

As a matter of fact, we have

$$\dot{z} = F_* \dot{x} = \frac{\partial}{\partial x} F(x)\dot{x} \tag{4.12}$$

and from step (iv) and (v),

$$g(z, u) = L_f z(x) = \frac{\partial}{\partial x} F(x) f(x, u). \tag{4.13}$$

It can be shown that $(\partial/\partial x)F(x)$ is nonsingular, so (4.12) and (4.13) imply

$$\dot{x} = f(x, u)$$

and (4.4a) gives

$$y = Cz = h(x).$$

This completes the proof of the theorem. □

Also, we see that there are some abundances in (4.3). We can drop the equations concerning a_{i0}, $i = 1, \cdots, m$, in (4.3). Let $\hat{b}_i$ be the corresponding vector functions obtained by deleting the $(\sigma_i + 1)$th entries, $i = 0, 1, \cdots, m-1$, and σ_0 is assumed to be zero. Then, by considering the integrability of the corresponding equations in almost the same fashion, we have Theorem 4.2.

THEOREM 4.2. *The system* (1.1) *admits an observer form* (1.6) *if and only if there exist m-tuple of integers* $(k_1, \cdots, k_m)$, $k_1 \geqq k_2 \geqq \cdots \geqq k_m > 0$, *and* $\sum_{i=1}^{m} k_i = n$, *such that we have the following*:

(1)′ *Conditions* (i) *and* (ii) *in Theorem* 2.3 *hold.*

(2)′ *There exist solutions* $g^1, \cdots, g^m$ *to* (2.26) *such that*

$$\text{(4.5)}' \qquad \operatorname{rank} \frac{\partial}{\partial x} \begin{bmatrix} \lambda_1(x, u) \\ \vdots \\ \lambda_m(x, u) \\ \hat{b}_1(x, u) \\ \vdots \\ \hat{b}_{m+p}(x, u) \\ h(x) \end{bmatrix} = m.$$

(3)′ *If we denote an* $(m+p)(n-m) \times (m+p)$ *matrix* $S'(x, u)$ *as*

$$\text{(4.6)}' \qquad S'(x, u) = \left[\left[\frac{\partial}{\partial x} \begin{bmatrix} \hat{b}_1(x, u) \\ \vdots \\ \hat{b}_{m+p}(x, u) \end{bmatrix} \left(\frac{\partial h}{\partial x} \right)^T \left[\frac{\partial h}{\partial x} \left(\frac{\partial h}{\partial x} \right)^T \right]^{-1} \middle| \frac{\partial}{\partial u} \begin{bmatrix} \hat{b}_1(x, u) \\ \vdots \\ \hat{b}_{m+p}(x, u) \end{bmatrix} \right] \right],$$

then $S'(x, u)$ *is block symmetric, i.e.,*

$$\text{(4.7)}' \qquad \overline{S'(x, u)} = (\overline{S'(x, u)})^T.$$

5. Examples.

Example 1. Consider the system

$$\begin{aligned} \dot{x}_{11} &= x_{12}, & & \\ \dot{x}_{12} &= x_{12}x_{21}, & y_1 &= x_{11}, \\ \dot{x}_{21} &= x_{12}. & y_2 &= x_{21}, \end{aligned}$$

Since

$$dh_1 = (1\ 0\ 0), \quad L_f(dh_1) = (0\ 1\ 0), \quad dh_2 = (0\ 0\ 1),$$

$$L_f(dh_2) = (0\ \ x_{21}\ \ x_{12}) = 0 \cdot dh_1 + x_{21}(L_f(dh_1)) + x_{12}(dh_2),$$

the coefficient before $L_f(dh_1)$ is $x_{21} \neq 0$, and so, by Theorem 2.3, it is not transformable to an observer form.

Example 2. Consider the system

$$\text{(5.1)} \qquad \begin{aligned} \dot{x}_{11} &= x_{12}, & & \\ \dot{x}_{12} &= x_{12}x_{21}, & y_1 &= x_{11}, \\ \dot{x}_{21} &= x_{11}. & y_2 &= x_{21}, \end{aligned}$$

This system is in special observable form, thus from the discussion of Example 7.4 of [6], it is transformable to an observer form. (This can also be seen in our next example.)

But system (5.1) does not satisfy (5.4) in [6]. Since, g^1, g^2 defined by (3.8), i.e., (5.3) in [6], are $g^1 = (0\ 1\ 0)^T$, and $g^2 = (0\ 0\ 1)^T$, and since

$$\mathrm{ad}_{(-f)}g^1 = \frac{\partial f}{\partial x} g^1 \frac{\partial g^1}{\partial x} f = \frac{\partial f}{\partial x} g^1 = \begin{bmatrix} 1 \\ x_{21} \\ 0 \end{bmatrix},$$

$$[\mathrm{ad}_{(-f)}g^1, g^2] = \begin{bmatrix} 0 \\ -1 \\ 0 \end{bmatrix} \neq 0.$$

That is, condition (5.4) in [6] fails.

Example 3. Consider system (5.1) again. Since

$$dh_1 = (1\ 0\ 0), \quad L_f(dh_1) = (0\ 1\ 0), \quad dh_2 = (0\ 0\ 1),$$
$$L_f(dh_2) = (1\ 0\ 0) = dh_1 + 0 \cdot L_f(dh_1) + 0 \cdot dh_2$$

with $k_1 = 2$, $k_2 = 1$, condition (i) of Theorem 3.1 is satisfied.

Moreover, consider (3.1). Obviously, $g^1 = (0\ 1\ 0)^T$. To determine g^2, we have only two equations:

$$\langle dh_1, g^2 \rangle = 0, \qquad \langle dh_2, g^2 \rangle = 1. \tag{5.2}$$

It can be easily seen that $g^2 = (0\ x_{11}\ 1)^T$ is a solution to these equations. And since

$$[g^1, g^2] = \frac{\partial g^2}{\partial x} g^1 - \frac{\partial g^1}{\partial x} g^2 = \frac{g^2}{x} g^1 = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0,$$

$$\mathrm{ad}_{(-f)}g^1 = \frac{\partial f}{\partial x} g^1 - \frac{\partial g^1}{\partial x} f = \frac{\partial f}{\partial x} g^1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & x_{21} & x_{12} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ x_{21} \\ 0 \end{bmatrix},$$

and thus

$$[\mathrm{ad}_{(-f)}g^1, g^2] = \frac{\partial g^2}{\partial x} \mathrm{ad}_{(-f)}g^1 - \frac{\partial}{\partial x} (\mathrm{ad}_{(-f)}g^1) g^2$$
$$= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ x_{21} \\ 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ x_{11} \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0;$$

similarly,

$$[\mathrm{ad}_{(-f)}g^1, g^1] = 0.$$

Therefore, (3.7) holds. This implies that the partial differential equations (3.2), which can now be explicitly written as

$$\frac{\partial \phi}{\partial z} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & x_{21} & x_{11} \\ 0 & 0 & 0 \end{bmatrix} \circ \phi(z), \tag{5.3}$$

are integrable. In fact,

$$\phi = \begin{bmatrix} z_{12} \\ z_{11} + z_{12} z_{21} \\ z_{21} \end{bmatrix}$$

is a solution to (5.3). Hence, under the coordinates transformation

$$x_{11}=z_{12},\quad x_{12}=z_{11}+z_{12}z_{21},\quad x_{21}=z_{21}$$

or

$$z_{11}=x_{12}-x_{11}x_{21},\quad z_{12}=x_{11},\quad z_{21}=x_{21}$$

system (4.1) is transformable into an observer form

$$\begin{aligned}&\dot{z}_{11}=-z_{12}^2=-y_1^2,\\&\dot{z}_{12}=z_{11}+z_{12}z_{21}=z_{11}+y_1y_2,\\&\dot{z}_{21}=z_{12}=y_1,\end{aligned}\qquad\begin{aligned}&y_1=z_{12},\\&y_2=z_{21}.\end{aligned}$$

By employing the computation procedure, we can obtain the state transformation much more simply. For this example, choose g^1 and g^2 as before, and now

$$Q(x)=(g^1\ \mathrm{ad}_{(-f)}g^1g^2)=\begin{bmatrix}0&1&0\\1&x_{21}&x_{11}\\0&0&1\end{bmatrix},$$

$$\left(\frac{\partial a}{\partial y_1},\frac{\partial a}{\partial y_2}\right)=\tilde{Q}^{-1}(x)(\mathrm{ad}_{(-f)}^2g^1,\mathrm{ad}_{(-f)}g^2)$$

$$=\begin{bmatrix}-2x_{11}&0\\x_{21}&x_{11}\\1&0\end{bmatrix}=\begin{bmatrix}-2y_1&0\\y_2&y_1\\1&0\end{bmatrix}.$$

Integrating the above equations, which are much more simple than (5.3), we have

$$a(y)=\begin{bmatrix}-y_1^2\\y_1y_2\\y_1\end{bmatrix},$$

and step (iv) gives the state transformation just obtained.

Example 4. Consider the system

$$\begin{aligned}&\dot{x}_1=x_1x_3-x_1x_4^2-x_2x_3^2+x_2x_3x_4^2-x_2x_4^4+(x_2^2+x_4)u+(x_3-x_4^2)u^2,\\&\dot{x}_2=x_1-x_2x_3+x_2x_4^2+u^2,\\&\dot{x}_3=2x_3x_4-2x_4^3+x_2u-x_4u^2,\qquad\dot{x}_4=x_3-x_4^2+u^2\\&y_1=x_2,\qquad y_2=x_4.\end{aligned}$$

Since

$$0(x,u)=\begin{bmatrix}0&1&0&0\\1&-x_3+x_4^2&-x_2&2x_2x_4\\0&0&0&1\\0&0&1&-2x_4\end{bmatrix}$$

is nonsingular, the system is observable with observability indices $k_1=k_2=2$. Now

$g^1(x, u) = (1\ 0\ 0\ 0)^T$, $g^2(x, u) = (x_2\ 0\ 1\ 0)^T$, and

$$\tilde{Q}(x, u) = (g^1\ \mathrm{ad}_{(-f)}g^1 g^2\ \mathrm{ad}_{(-f)}g^2)$$

$$= \begin{bmatrix} 1 & x_3 - x_4^2 & x_2 & -x_2 x_4^2 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 2x_4 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\left(\frac{\partial a}{\partial y_1}, \frac{\partial a}{\partial y_2}\right) = \tilde{Q}^{-1}(x, u)(\mathrm{ad}^2_{(-f)}g^1, \mathrm{ad}^2_{(-f)}g^2) = (b_1(x, u), b_2(x, u))$$

$$= \begin{bmatrix} 0 & u \\ 0 & 0 \\ u & 0 \\ 0 & 0 \end{bmatrix}$$

and $b_3(x, u)$ computed by (4.1) is $b_3(x, u) = (x_4\ 2u\ x_2\ 2u)^T$. It is easily seen that b_1, b_2, b_3 are functions of y_1, y_2, and u. Now the matrix $S(x, u)$ is

$$S(x, u) = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 0 & 2 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{array}\right].$$

Obviously, $S(x, u)$ is block symmetric. Finally, the computation procedure yields

$$a(y, u) = (y_2 u\ \ u^2\ \ y_1 u\ \ u^2)^T$$

and the state transformation

$$z(x) = F(x)$$

$$= \begin{bmatrix} x_1 - x_2 x_3 + x_2 x_4^2 \\ x_2 \\ x_3 - x_4^2 \\ x_4 \end{bmatrix}.$$

The observer form is

$$\dot{z}_1 = y_2 u, \qquad \dot{z}_2 = z_1 + u^2,$$

$$\dot{z}_3 = y_1 u, \qquad \dot{z}_4 = z_3 + u^2.$$

6. Conclusions. In this paper, we have identified a class of nonlinear systems whose observer design problem can be solved by the observer error linearization

approach. We have considered systems both with and without inputs. For systems without inputs, we have obtained a set of necessary and sufficient conditions in terms of the Lie algebra, which is a correction to a theorem in [6]. Moreover, we have proposed a computation procedure for the practical computation of the state transformation, and on the basis of this computation procedure we have derived a different set of necessary and sufficient conditions described in rank conditions of matrices. It is noted that our approach in dealing with systems with inputs is different from that of Krener and Respondek in [6]. Also, it may be of interest to consider output coordinates change that will certainly enlarge the class of nonlinear systems. For this point, we refer the reader to the paper mentioned above.

However, it should be noted that the conditions given here depend on the choice of solutions to the linear algebraic equations (2.26). How to choose these solutions remains open. But, as have been seen, the conditions are useful in understanding the observer error linearization problem and in giving various sufficient conditions.

REFERENCES

[1] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47–52.

[2] D. BESTLE AND M. ZEITZ, *Canonical form observer design for nonlinear time variable systems*, Internat. J. Control, 38 (1983), pp. 419–431.

[3] B. JAKUBCZYK AND W. RESPONDEK, *On the linearization of control systems*, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astronom. Phys., 28 (1980), pp. 517–522.

[4] L. R. HUNT AND R. SU, *Linear equivalents of nonlinear time varying systems*, in International Symposium on the Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 119–123.

[5] C. W. LI AND L. W. TAO, *Observing non-linear time-variable systems through a canonical form observer*, Internat. J. Control, 44 (1986), pp. 1703–1713.

[6] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.

[7] X. H. XIA AND W. B. GAO, *Nonlinear observer design by observer canonical forms*, Internat. J. Control, 47 (1988), pp. 1081–1100.

[8] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry*, Publish or Perish, Berkeley, CA, 1979.

[9] J. P. GAUTHIER AND G. BORNARD, *Observability for any $u(t)$ of a class of nonlinear systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 922–926.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 1, pp. 217-234, January 1989

012

REACHABILITY OF A CLASS OF INFINITE-DIMENSIONAL LINEAR SYSTEMS: AN EXTERNAL APPROACH WITH APPLICATIONS TO GENERAL NEUTRAL SYSTEMS*

YUTAKA YAMAMOTO†

Abstract. This paper studies the question of quasi- (approximate) reachability of the standard observable realizations of pseudorational impulse responses introduced by the author. The framework places the current theory of retarded and neutral delay-differential systems into a unified input/output framework. Several necessary and sufficient conditions for quasi-reachability are derived. In particular, new criteria for quasi-reachability and eigenfunction completeness are obtained for general delay-differential systems with no restriction on the type of delays. Furthermore, as a byproduct, the theory leads to necessary and sufficient conditions for approximate left coprimeness of matrices with distribution entries. Examples are discussed to illustrate the theory.

Key words. pseudo-rationality, infinite-dimensional systems, reachability, eigenfunction completeness, left coprimeness, distributions, delay-differential systems

AMS(MOS) subject classifications. 93B05, 93C20

1. Introduction.

Consider the following delay-differential system of neutral type:

$$\frac{d}{dt}x(t) = F_0x(t) + \sum_{i=1}^{N} F_ix(t-h_i) + \sum_{i=1}^{N} F_{-i}\dot{x}(t-h_i) + \int_{-h}^{0} E(\tau)x(t+\tau)\,d\tau + Gu(t), \tag{1.1}$$

where $0 < h_1 < h_2 < \cdots < h_N = h$. Approximate controllability of this system has been investigated by a number of authors ([6], [12], [13], [18]). One of them, Manitius [14], [15], gave a complete rank condition for the retarded case (i.e., $F_{-i} = 0$, $i = 1, \cdots, N$). O'Connor and Tarn [17] extended his results to the neutral case, but gave a complete algebraic condition only for the case $N = 1$ with no distributed delays. Salamon [20] extended these results to the case with delays in input, as well as giving the discussion of the general case using the notion of small solutions. However, a concrete algebraic criterion has not been obtained for the case when there is any distributed delay; obtaining such a criterion in the situation with general delays is left as an open problem. In view of the complexity of formulas there, it appears to be difficult to obtain a generalization in this setting.

These approaches all associate an a priori chosen function state space model (e.g., M_2- or W_2^1-) to (1.1) and then discuss approximate function space controllability (reachability) there. There are, however, cases in which a priori association of a state space is either inappropriate or inconvenient. For example, in the study of a servo problem, we often need to consider a reference signal generator, which is specified by the transfer function, and this transfer function does not easily fall into the existing category of function space models. The recently introduced control scheme called *repetitive control* [8], [9], [16], which uses a model $1/(\exp(Ls) - 1)$, is a typical example of such a case. Another example where a priori association of a state space causes difficulty is the case of (1.1), in which there is a pole-zero cancellation. In such a case,

* Received by the editors May 5, 1986; accepted for publication (in revised form) March 14, 1988. This work was supported in part by Scientific Research Grant-in-Aid 60750375 from the Ministry of Education, Science and Culture, Japan.

† Department of Applied Systems Science, Faculty of Engineering, Kyoto University, Kyoto 606, Japan.

although we may expect to obtain an irredundant model after pole-zero reduction, the reduced state space will not be M_2 or W_2^1, because these spaces are not closed under such an operation.

In view of these, it is also desirable to develop an approach more directly associated with transfer functions, impulse responses, and realization theory, so that the analysis is not affected by the presumptive choice of a state space.

In finite-dimensional systems, it is well known that reachability (controllability) can be studied from both internal and external points of view. It is also known that the latter is advantageous in relating reachability to irredundancy of the external representation. For example, given a fractional representation $Q^{-1}(s)P(s)$ (Q, P = polynomial matrices) of a transfer function, we can always associate with it a standard observable realization Σ^Q (known as the *Fuhrmann* realization), and this realization is reachable if and only if the matrices Q and P are left coprime [5]. It is, therefore, natural to attempt to generalize this fact to infinite-dimensional systems. However, in spite of a number of investigations on fractional representations of irrational transfer matrices (e.g., [2], [3], [10], [25]), the reachability question above has not been studied in connection with the external behavior.

It is then natural to ask the following questions in this context:

(i) Given a transfer matrix of a delay-differential equation, what can be said about its realization? Is there an analogue of the Fuhrmann realization in this context?

(ii) If there is, what can be said about its reachability in terms of the transfer matrix (or its fractional representation)?

The first question has been studied in [28], and it has been shown that a precise analogue of the Fuhrmann observable realization exists. This has been done by introducing a class of fractional representation called *pseudo-rational.* At least all transfer matrices of delay-differential systems belong to this class. (In some other frameworks (e.g., [2], [3]) using the algebra $\mathcal{A}$ of stable impulse responses, this is not true.) Concerning (ii), it is also shown in [28] that the obtained observable realization Σ^Q (see § 2 for the definition) is approximately controllable (from here on we use the term *quasi-reachable* for consistency with [28]) if and only if the associated fractional representation of the transfer matrix is left coprime in some weaker sense. However, this condition gives only an abstract condition, and its consequence on a concrete reachability criterion has not been fully investigated there.

In this paper, we shall derive necessary and sufficient conditions for this reachability and coprimeness according to the above program. The results, when specialized to delay-differential systems, generalize the existing ones to those for systems with distributed/noncommensurable delays. To see the basic idea of the method, let us first write the transfer function of (1.1) in the folllowing form, by supplementing a fictitious output equation $y(t) := x(t-h)$:

$$\begin{aligned} &W(s) = \hat{Q}(s)^{-1}\hat{P}(s), \\ &\hat{Q} := [e^{-hs}(sI - F_0) - \sum e^{(-h+h_i)s}(F_i - sF_{-i}) - \hat{E}(s)], \\ &\hat{P} := G. \end{aligned} \tag{1.2}$$

Their inverse Laplace transforms induce the following representation for the impulse response matrix A:

$$\begin{aligned} &A = Q^{-1} * P, \\ &Q := \delta_{-h}(\delta' I - F_0) - \sum \delta_{-h+h_i}(F_i - \delta' F_{-i}) - E, \\ &P := G\delta, \end{aligned} \tag{1.3}$$

where $\delta :=$ Dirac distribution at 0, $\delta' :=$ its derivative, $\delta_{-a} :=$ Dirac distribution at $-a$. These entities are most naturally investigated in the convolution algebra $\mathscr{E}'(\mathbf{R}^-)$, which is the space of Schwartz distributions having compact support contained in $(-\infty, 0]$ (see the end of this section for the definition).

By using the convolution algebra structure of $\mathscr{E}'(\mathbf{R}^-)$, we will prove the following:

(i) The Fuhrmann-like observable realization Σ^Q is spectrally complete (i.e., the space spanned by eigenfunctions is dense in the state space) if and only if the matrix Q assumes full rank near the origin (see § 3 for a precise statement).

(ii) Σ^Q is quasi-reachable if and only if it is spectrally reachable (i.e., any eigenspace is reachable) and the matrix $[Q, P]$ assumes full rank near the origin.

In the above results, Q admits any type of delays (distributed/noncommensurable). Therefore, they generalize the existing criteria by Manitius [14], O'Connor and Tarn [17], and Salamon [20] to the general case *with no restrictions on the type of delays.*

The paper is organized as follows. In § 2 we review the basic realization framework given in [27]–[29], especially a type of fractional representation and its associated observable realization Σ^Q. Section 3 gives a general necessary and sufficient condition for eigenfunction completeness. In § 4, this result is applied to derive conditions for approximately controllability. Section 5 gives an application to delay-differential systems along with the discussion of an example. It is seen that the obtained criterion is simpler than the discussion involving small solutions.

Notation and conventions. In what follows, the following notation will be used. All vector spaces and function spaces are considered over a fixed field k, which is either $\mathbf{R}$ or $\mathbf{C}$. Functions and distributions are also k-valued. As usual, for a set V, V^n denotes the set of n-fold direct product of V. For a vector $x \in V^n$, x^T denotes its transpose. If V is a topological space, we endow V^n with the usual product topology. For a ring (or algebra) R, $R^{p\times m}$ denotes the set of $p \times m$ matrices with entries in R. When R is also a topological space, $R^{p\times m}$ is understood to be endowed with the product topology as above. To simplify notation we may sometimes drop these superscripts; for example, when it is clear that $x(t)$ is an n-vector, *we may write* $x(t) \in L^2[0, T]$, *instead of writing* $x(t) \in (L^2[0, T])^n$. We shall always consider the bilateral Laplace transform and regard the one-sided Laplace transform as a special case of it. The Laplace transform of a distribution α (if it exists) will be denoted by $\hat{\alpha}(s)$.

We assume standard knowledge on distribution theory, such as can be found in Schwartz's account [23]. Some familiarity with basic notions of the theory of locally convex topological vector spaces is also assumed [21], [23]. However, some of the following function spaces, although quite fundamental, may not be encountered in standard textbooks (see [26], [28] for details).

$L^2[a, b] :=$ the space of Lebesgue square-integrable functions on $[a, b]$.

$L^2_{\mathrm{loc}}[0, \infty) :=$ the space of functions square-integrable on every bounded interval.

$\Omega := \cup_{n=1}^{\infty} L^2[-n, 0]$ as a set. This space is endowed with the natural inductive limit topology [21], [26] induced by the sequence of spaces $\{L^2[-n, 0]\}_{n=1}^{\infty}$.

$\mathscr{D}(\mathbf{R}) :=$ the space of C^∞-functions defined on $\mathbf{R}$ with the usual topology (Schwartz [23]).

$\mathscr{D}'(\mathbf{R}) :=$ the space of distributions on $\mathbf{R}$. The support of a distribution α will be denoted by $\operatorname{supp} \alpha$. When $\alpha \in (\mathscr{D}'(\mathbf{R}))^n$, $\operatorname{supp} \alpha$ is understood to be the union of the supports of its entries.

$\mathscr{D}[0, \infty) :=$ the space of C^∞-functions *defined on* $[-\infty, \infty)$ *having compact support contained in* $[0, \infty)$.

$\mathscr{D}(-\infty, 0] :=$ the space of C^∞-functions *defined on* $(-\infty, \infty)$ *having compact support contained in* $(-\infty, 0]$.

$\mathscr{D}'[0, \infty) :=$ the dual space of $\mathscr{D}[0, \infty)$.

$\mathscr{D}'_+ :=$ the space of distributions on $\mathbf{R}$ *with support bounded on the left.* This is an algebra with respect to convolution with idenity δ (the Dirac delta distribution at the origin). The delta distribution at point a will be denoted by δ_a, and its derivative will be denoted by δ'_a. Convolution will be denoted by $*$, as usual.

$\mathscr{D}'_- :=$ the space of distributions on $\mathbf{R}$ with support bounded on the right. This is also a convolution algebra with identity δ.

$\mathscr{E}'(\mathbf{R}) :=$ the space of distributions on $\mathbf{R}$ with *compact* support.

$\mathscr{E}'(\mathbf{R}^-) :=$ the space of distributions on $\mathbf{R}$ having *compact support contained in* $(-\infty, 0]$.

Throughout the paper, duality will be denoted by $\langle \cdot, \cdot \rangle$, that is, the value of a distribution α evaluated at a test function φ will be denoted by $\langle \alpha, \varphi \rangle$ or by $\langle \varphi, \alpha \rangle$. Among the spaces above the following inclusions hold:

$$\mathscr{D}(-\infty, 0] \subset \Omega \subset \mathscr{E}'(\mathbf{R}^-) \subset \mathscr{E}'(\mathbf{R}) \subset \mathscr{D}'_+, \qquad \mathscr{E}'(\mathbf{R}^-) \subset \mathscr{E}'(\mathbf{R}) \subset \mathscr{D}'_-.$$

$\mathscr{D}(-\infty, 0]$ is dense in Ω, and Ω is dense in $\mathscr{E}'(\mathbf{R}^-)$. The space $\mathscr{D}'[0, \infty)$ is *not* a subspace of $\mathscr{D}'_+$, but there exists a *surjective* (continuous, of course) projection $\pi : \mathscr{D}'_+ \to \mathscr{D}'[0, \infty)$ which is induced from the obvious canonical inclusion $j : \mathscr{D}[0, \infty) \to \mathscr{D}_-$, that is,

$$\langle \pi\alpha, \varphi \rangle := \langle \alpha, j\varphi \rangle. \tag{1.7}$$

2. Preliminaries. Consider the usual linear (zero-initial state) input/output correspondence:

$$f(\omega)(t) = \int_0^t A(t-\tau)\omega(\tau)\, d\tau, \tag{2.1}$$

where ω is an input and A is an impulse response matrix that does not necessarily induce a finite-dimensional realization. For the purpose of realization theory, it is convenient to convert (2.1) to another form. In view of shift-invariance and causality of (2.1), it can be written as

$$f(\omega)(t) = \int_{-\infty}^0 A(t-\tau)\omega(\tau)\, d\tau \tag{2.2}$$

if the input ω is applied on $(-\infty, 0]$ and has compact support. Since the system must be causal, observation of all $f(\omega)(t)$ for all $t \geqq 0$ must be enough to determine the internal state space structure from (2.2) [27]. We thus take the *input function space* to be Ω^m ($m :=$ number of input channels), and the *output function space* Γ^p ($p :=$ number of output channels), so that inputs are applied before time zero and outputs are observed after zero. These spaces are naturally equipped with the left shift operators $\{\sigma_t\}$, $t \geqq 0$:

$$\begin{aligned} (\sigma_t\omega)(s) &:= \begin{cases} \omega(s-t) & \text{for } s < -t, \\ 0 & \text{for } 0 \leqq s \leqq 0, \end{cases} \quad \omega \in \Omega^m, \\ (\sigma_t\gamma)(s) &:= \gamma(s-t), \quad \gamma \in \Gamma^p. \end{aligned} \tag{2.3}$$

Thus we define a (constant, linear) *input/outmap* f to be a continuous linear map that commutes with these left shifts [27]. It is known [26], [27] that f can then be represented by (2.2), where A is a $p \times m$ matrix (Radon) measure on $[0, \infty)$.

Now note that (2.2) can be written as

$$f(\omega) = \pi(A * \omega), \qquad w \in \Omega^m, \tag{2.4}$$

where π is the truncation operator defined by (1.7). The matrix measure A here is called the *impulse response matrix* of the input/output map f. Let us introduce a fractional representation for such impulse response matrices.

DEFINITIONS 2.5. Let Q be a $p \times p$ matrix with entries in $\mathscr{E}'(\mathbf{R}^-)$. Q is said to be of *normal type* if the following conditions are satisfied:

(i) Q is invertible over $\mathscr{D}'_+$ with respect to convolution, i.e., det Q (computed in terms of convolution) is invertible with respect to convolution;

(ii) $\operatorname{ord}(\det Q)^{-1} = -\operatorname{ord}(\det Q)$,

where ord α denotes the *order* of a distribution α [23]. An impulse response matrix A is said to be *pseudo-rational* if it can be written in the form

$$A = \pi(Q^{-1} * P) \tag{2.6}$$

for some matrices Q and P over $\mathscr{E}'(\mathbf{R}^-)$, where the $p \times p$ matrix Q is of normal type.

Although not all impulse response matrices are pseudo-rational, at least all delay-differential impulse responses are known to be pseudo-rational [29], [30]. As can be seen from (1.2) or (1.3), there will be *no difficulty* in dealing with the distributed delay case.

We now define the standard observable realization Σ^Q associated with an impulse response matrix $A = Q^{-1} * P$. Let X^Q be a subspace of Γ^p defined by

$$X^Q := \{x(t) \in \Gamma^p;\ \pi(Q * x) = 0\}. \tag{2.7}$$

The condition $\pi(Q * x) = 0$ is clearly equivalent to $\operatorname{supp}(Q * x) \subset (-\infty, 0]$. By the separate continuity of convolution, X^Q is a closed subspace of Γ^p. The family $\{\sigma_t\}$ of left-shift operators constitutes a strongly continuous semigroup in Γ^p, and X^Q is easily seen to be a σ_t-invariant subspace of Γ^p. Let F denote the infinitesimal generator of this semigroup σ_t in X^Q; this is nothing but the differential operator d/dt. Define Σ^Q as follows:

(2.8) State space $= X^Q$;

State transition equation:

(2.9) $$\varphi(t, x, u) := \sigma_t x + \pi(A * \sigma_t^l u), \quad x \in X^Q, \quad (\sigma_t^l u)(s) := u(s+t),$$

where the right-hand side is the state at time t resulting from input u and initial state x;

Output equation:

(2.10) $$y = Hx := x(0).$$

This definition yields the mappings $g: \Omega^m \to X^Q$ and $h: X^Q \to \Gamma^p$ as follows:

$$g(\omega) := \varphi(T, 0, \sigma_T^r \omega), \tag{2.11}$$

$$h(x)(t) := H\sigma_t x, \qquad t \geqq 0. \tag{2.12}$$

Here T is any positive number such that $[-T, 0] \supset \operatorname{supp} \omega$ and $(\sigma_T^r \omega)(t) := \omega(t - T)$. It is easy to see that (2.11) is independent of the choice of T. We say that Σ^Q *realizes* an input/output map f (or its impulse response A) if $f = hg$; it is *quasi-reachable* (commonly referred to as approximately controllable) if the reachable set $g(\Omega^m)$ is dense in X^Q; *topologically observable* if h is a topological isomorphism; and *topologically observable in bounded time* if X^Q is topologically isomorphic to $h(X^Q)|_{[0,T]}$ for some $T > 0$.

Σ^Q is easily seen to give a realization of A. Furthermore, it is known that this realization possesses various desirable properties.

THEOREM 2.13 [27]–[29]. (i) Σ^Q *is always topologically observable in bounded time. This means that the initial state determination is well posed, and can be done based on the output data* $y(t)|_{[0,T]}$ *on a uniformly bounded time interval* $[0, T]$.

(ii) Σ^Q *is quasi-reachable if and only if the matrices* Q *and* P *are approximately left coprime, i.e.,*

$$Q * R_n + P * S_n \to \delta I_p \tag{2.14}$$

for some sequences R_n *and* S_n *of matrices with entries in* $\mathscr{E}'(\mathbf{R}^-)$.

We shall investigate the quasi-reachability of Σ^Q using (ii). To this end, we shall need some technical lemmas which we now summarize.

For a distribution $\psi \in \mathscr{D}'_-$, define $r(\psi)$ to be the *supremum* of the support of ψ, i.e.,

$$r(\psi) := \sup\{t \in \operatorname{supp} \varphi\}. \tag{2.15}$$

The following lemma is a consequence of the well-known theorem of Titchmarsh on convolution [4, p. 224].

LEMMA 2.16. *For* $\varphi, \psi \in \mathscr{D}'_-$,

$$r(\varphi * \psi) = r(\varphi) + r(\psi).$$

In § 3, we need to use a representation of the dual space of X^Q. To this end, we first introduce the following duality between Ω^p and Γ^p:

$$(\omega, \gamma) := \int_{-\infty}^{0} \omega^T(\tau)\gamma(-\tau)\, d\tau = \int_0^{\infty} \omega^T(-\tau)\gamma(\tau)\, d\tau = (\omega^T * \gamma)(0). \tag{2.17}$$

Then Ω and Γ turn out to be topologically dual to each other [29]. (Needless to say, $(\Omega^p)' \cong \Gamma^p$.) With respect to this duality, we have the following lemma.

LEMMA 2.18. *Let* X^Q *be as above, where* Q *is of normal type. Define the polar (orthogonal complement) of* X^Q *by*

$$(X^Q)^0 := \{\omega \in \Omega^p;\ (\omega, \gamma) = 0 \text{ for all } \gamma \in X^Q\}.$$

Then,

$$\begin{aligned}(X^Q)^0 &= \{\omega \in \Omega^p;\ \omega^T * Q^{-1} \in \mathscr{E}'(\mathbf{R}^-)\} \\ &= \overline{\{Q^T * \psi;\ \psi \in (\mathscr{D}(-\infty, 0])^p\}}.\end{aligned} \tag{2.19}$$

Proof. Suppose that ψ belongs to the right-hand side of (2.16) and γ belongs to X^Q. Observe that

$$(Q^T * \psi, \gamma) = (\psi^T * Q * \gamma)(0). \tag{2.20}$$

$Q * \gamma$ has compact support contained in $(-\infty, 0]$ because γ belongs to X^Q. Then the right-hand side of (2.20) is clearly zero, which implies that the left-hand side of (2.19) contains the right-hand side.

Conversely, suppose that $\omega \in \Omega^p$ belongs to $(X^Q)^0$. Suppose first that ω belongs to $\mathscr{D}(-\infty, 0]$. Take any $\varphi \in (\mathscr{D}(-\infty, 0])^p$. Then $\gamma := \pi(Q^{-1} * \varphi) \in X^Q$. Since it is easy to see that $\pi(\omega^T * \pi(Q^{-1} * \varphi)) = \pi(\omega^T * Q^{-1} * \varphi)$, we have

$$\begin{aligned}\pi(\omega^T * Q^{-1} * \varphi)(0) &= \pi(\omega^T * \pi(Q^{-1} * \varphi))(0) = (\omega^T * \pi(Q^{-1} * \varphi))(0) \\ &= (\omega, \pi(Q^{-1} * \varphi)) = 0\end{aligned}$$

because $\omega \in (X^Q)^0$. Since this is true for any $\delta_{-t} * \varphi$ $(t > 0)$, we have $\pi(\omega^T * Q^{-1} * \varphi) = 0$, and hence $\omega^T * Q^{-1} * \varphi \in (\mathscr{E}'(\mathbf{R}^-))^p$. Since this holds for any $\varphi \in (\mathscr{D}'(-\infty, 0])^p$, we

must have $\omega^T * Q^{-1} \in (\mathscr{E}'(\mathbf{R}^-))^p$. Thus $\omega = Q^T * \psi$ for some $\psi \in (\mathscr{E}'(\mathbf{R}^-))^p$. Since $\mathscr{D}(-\infty, 0]$ is dense in Ω, this implies that ω belongs to the second term of the right-hand side of (2.19). Then by continuity, we see that this is true for any $\omega \in X^Q$. □

We denote the right-hand side of (2.19) *by* ${}_QX$. Then we have the following lemma characterizing the dual space of X^Q.

LEMMA 2.21. *The dual space* $(X^Q)'$ *of* X^Q *is topologically isomorphic to* $\Omega^p/{}_QX$.

Proof. The proof is immediate from Lemma 2.18 and Schaefer [21, Chap. 4, § 4]. □

Before closing this section, we give a remark on another "realization" obtained from Σ^Q. Introduce the graph norm to the domain $D(F)$ of the infinitesimal generator F of the shift semigroup σ_t, i.e.,

$$\|x\|_F := \|x\| + \|Fx\|,$$

where $\|x\|$ is the norm in X^Q. Restricting σ_t to $D(F)$, we get another strongly continuous semigroup. Let us restrict the class of input and output functions to the spaces analogous to Ω^m and Γ^p that locally belong to the class of the first-order Sobolev space W_2^1. Then the state transition and the output equation given by (2.9) and (2.10) are well defined, and we obtain a well-defined system with state space $D(F)$ (although with different input/output spaces). Denote this system by Σ_F^Q. The following proposition, which claims the equivalence of quasi-reachability of Σ^Q and Σ_F^Q, will become necessary in § 5 for discussing the M_2-reachability and W_2^1-reachability for delay-differential systems.

PROPOSITION 2.22. Σ^Q *is quasi-reachable if and only if* Σ_F^Q *is also.*

Proof. See Appendix A for the proof.

3. Eigenfunction completeness. Let Σ^Q be the system defined in the previous section. We shall investigate the completeness of eigenfunctions of this system. As noted in § 2, the results below apply to neutral systems with *distributed delays* as well as those with *noncommensurable delays.* Let us first note that this system shares some nice spectral properties of delay-differential systems.

THEOREM 3.1. *Let* Σ^Q *be the system given by* (2.9), (2.10). *Let* $\sigma(F)$ *denote the spectrum of the infinitesimal generator* F *of the transition semigroup* σ_t. *Then the following statements hold*:

(i) $\sigma(F) = \{\lambda \in \mathbf{C};\ \det \hat{Q}(\lambda) = 0\}$;

(ii) *Every* $\lambda \in \sigma(F)$ *is an eigenvalue, and the generalized eigenspace* M_λ, *corresponding to* λ, *has finite dimension.*

(iii) $\dim M_\lambda = m\,(\det \hat{Q}(s), \lambda)$, *where the right-hand side denotes the multiplicity of* λ *as a zero of* $\det \hat{Q}(s)$.

(iv) *A vector* v *belongs to* M_λ *if and only if it is expressible in the form*

$$v = \sum_{j=0}^{m-1} \left(\frac{t^j e^{\lambda t}}{j!} \right) a_{j+1},$$

where

$$\begin{bmatrix} \hat{Q}(\lambda) & \cdots & & \hat{Q}^{(m-1)}(\lambda)/(m-1)! \\ & \ddots & & \vdots \\ 0 & & \hat{Q}(\lambda) & \hat{Q}^{(1)}(\lambda) \\ & & & \hat{Q}(\lambda) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_{m-1} \end{bmatrix} = 0.$$

Here $\hat{Q}^{(i)}$ *is the ith derivative of* $\hat{Q}$, *and* m *is the multiplicity of* λ *as a zero of* $\det \hat{Q}(s)$.

For a proof, see [28]. The last statement is not given in [28], but it is obtained via minor modification of the proof given in Hale [7].

Let $M := \operatorname{span}\{M_\lambda; \lambda \in \sigma(F)\}$, i.e., the space spanned by all generalized eigenspaces. Let us prepare the following terminology.

DEFINITION 3.2. The system Σ^Q is said to be *spectrally complete* if the above-defined space M is dense in X^Q.

We raise the following question. *Under what condition is Σ^Q spectrally complete?* In other words, when is the set of all eigenfunctions complete in X^Q?

The same problem has been studied for retarded delay-differential systems by Manitius [13], [14] and by O'Connor and Tarn [17] for neutral systems. In the latter treatment, a concrete condition is given only for the case of one point delay. We shall generalize the results of [17] to Σ^Q. When specialized to delay-differential systems, this means that there is *no restriction* on the type of delays involved. As we shall see, the generalization here clearly exhibits the fact that spectral completeness depends only on *the local behavior of Q near the origin.*

The following lemma is an immediate consequence of Lemma 2.21, which asserts that the zero element of the dual space of X^Q is precisely ${}_QX$:

LEMMA 3.3. *M is dense in X^Q if and only if the following statement is true: If $\varphi \in \Omega^p$ satisfies $(\varphi, \varphi_k) = 0$ for all $\varphi_k \in M$, then φ belongs to ${}_QX$.*

The following proposition gives a characterization for the statement $(\varphi, \varphi_k) = 0$ for all $\varphi_k \in M$:

PROPOSITION 3.4. *An element $\varphi \in \Omega^p$ satisfies $(\varphi, \varphi_k) = 0$ for all φ_k in M if and only if $\hat{\varphi}(s)^T \hat{Q}(s)^{-1}$ is an entire function of s.*

Since we have Theorem 3.1, the proof is entirely similar to that for the corresponding result of O'Connor and Tarn [17], and hence is omitted.

The following stronger version of Lemma 3.3 is also an easy consequence of the separate continuity of the bilinear functional $\langle \cdot, \cdot \rangle$, the denseness of $\Omega \subset \mathscr{E}'(\mathbf{R}^-)$ and Proposition 3.4.

PROPOSITION 3.5. *Σ^Q is spectrally complete if and only if $\hat{\varphi}(s)^T \hat{Q}^{-1}(s) = \text{entire}$, $\varphi \in (\mathscr{E}'(\mathbf{R}^-))^p$, implies $\varphi^T * Q^{-1} \in (\mathscr{E}'(\mathbf{R}^-))^p$.*

Proof. Necessity. Take any $\varphi \in (\mathscr{E}'(\mathbf{R}^-))^p$ such that $\hat{\varphi}(s)^T \hat{Q}^{-1}(s)$ is entire. Take a sequence $\{\rho_n\} \in \mathscr{D}(-\infty, 0]$ such that

$$\operatorname{supp} \rho_n \subset [-1/n, 0]; \tag{3.6}$$

$$\int \rho_n(t)\, dt = 1. \tag{3.7}$$

Then it is well known (e.g., Schwartz [23]) that (i) $\rho_n * \varphi \to \varphi$, and (ii) $\rho_n * \varphi \in \Omega^p$. It follows that $\rho_n * \varphi$ satisfies the same hypothesis as φ, so that $\rho_n * \varphi$ must belong to $(X^Q)^0 = {}_QX$. Hence $\rho_n * \varphi^T * Q^{-1}$ all belong to $(\mathscr{E}'(\mathbf{R}^-))^p$ by Lemma 2.18. Since $\{\rho_n * \varphi^T * Q^{-1}\}$ is convergent, it follows that $\varphi^T * Q^{-1}$ belongs to $(\mathscr{E}'(\mathbf{R}^-))^p$.

Sufficiency. Conversely, suppose that $\varphi \in \Omega^p$ satisfies $\hat{\varphi}(s)^T \hat{Q}^{-1} = \text{entire}$. Then we have $\varphi^T * Q^{-1} = \omega^T \in (\mathscr{E}'(\mathbf{R}^-))^p$ by hypothesis. Since there exists a sequence $\{\omega_n\} \subset \Omega^p$ that converges to ω, we have $\varphi^T = \lim \omega_n^T * Q$, which implies $\varphi \in {}_QX$ by Lemma 2.18. Hence, by Lemma 3.3, Σ^Q is spectrally complete. □

The proof of the following key lemma will be given in Appendix B.

LEMMA 3.8. *Let $\varphi \in (\mathscr{E}'(\mathbf{R}^-))^p$ and suppose that $\hat{\varphi}(s)^T \hat{Q}^{-1}(s)$ is an entire function. Then $\varphi^T * Q^{-1}$ belongs to $(\mathscr{E}'(\mathbf{R}))^p$, i.e., each of its entries is a distribution with compact support (which is not necessarily contained in $(-\infty, 0]$).*

This lemma shows that if φ is orthogonal to M, then it is always expressible as $\varphi = Q^T * \psi$ for some ψ in $(\mathscr{E}'(\mathbf{R}))^p$. However, this ψ does not necessarily belong to $(\mathscr{E}'(\mathbf{R}^-))^p$, i.e., its support may not be contained in $(-\infty, 0]$. Therefore, the condition of Proposition 3.5 may not be satisfied. To obtain a condition for spectral completeness,

we need only to exclude the possibility of $\varphi \in (\mathscr{E}'(\mathbf{R}^-))^p$, but $\psi \notin (\mathscr{E}'(\mathbf{R}^-))^p$. This leads to the following theorem.

THEOREM 3.9. *The system Σ^Q is spectrally complete if and only if there exists no $\psi \in (\mathscr{E}'(\mathbf{R}))^p$ such that $r(\psi) > 0$ and $r(\psi^T * Q) \leqq 0$.*

Proof. In view of the definition of $r(\psi)$, $\psi \in (\mathscr{E}'(\mathbf{R}))^p$ belongs to $(\mathscr{E}'(\mathbf{R}^-))^p$ if and only if $r(\psi) \leqq 0$. Then the theorem is immediate from Proposition 3.5 and Lemma 3.8. □

To give a more concise expression to Theorem 3.9, let us prepare the following algebraic notions.

LEMMA 3.10. *Let $J := \{\varphi \in \mathscr{E}'(\mathbf{R}^-);\ r(\varphi) < 0\}$. Then J is a prime ideal of $\mathscr{E}'(\mathbf{R}^-)$.*

Proof. That J is closed under addition is obvious. Suppose $\varphi \in J$ and $a \in \mathscr{E}'(\mathbf{R}^-)$. Then we have

$$r(a * \varphi) = r(a) + r(\varphi) \leqq r(\varphi) < 0,$$

by Lemma 2.16. Hence $a * \varphi \in J$, so that J is an ideal.

Now suppose that $a * b \in J$, i.e., $r(a * b) < 0$. Since

$$r(a * b) = r(a) + r(b)$$

by Lemma 2.16, either $r(a) < 0$ or $r(b) < 0$. That is, $a \in J$ or $b \in J$. Thus J is a prime ideal. □

Lemma 3.10 enables us to form the quotient ring (algebra) $\mathscr{A} := \mathscr{E}'(\mathbf{R}^-)/J$, and this ring is an *integral domain* (i.e., it has no zero divisors) because J is a prime ideal. Therefore, we can further construct its quotient field $\mathscr{F}$ (the field of fractions formed by elements of $\mathscr{A}$). Let $\theta : \mathscr{E}'(\mathbf{R}^-) \to \mathscr{F}$ denote the composition of the canonical projection: $\mathscr{E}'(\mathbf{R}^-) \to \mathscr{A} = \mathscr{E}'(\mathbf{R}^-)/J$ with the inclusion: $\mathscr{A} \to \mathscr{F}$. In what follows, when we speak of the *rank of a matrix $W \in (\mathscr{E}'(\mathbf{R}^-))^{p \times m}$ over $\mathscr{F}$, we shall always mean the rank of the matrix $\theta(W)$ considered over $\mathscr{F}$.* Observe that *an element $w \in (\mathscr{E}'(\mathbf{R}^-))^p$ is nonzero when considered over $\mathscr{A}$* (or $\mathscr{F}$) *if and only if $r(w) = 0$.*

We are now ready to state and prove the main result of this section, which is a generalization of the existing results in [14], [17], [20]. Note that no restriction on the type of delays is imposed.

THEOREM 3.11. *The system Σ^Q is spectrally complete if and only if*

$$\operatorname{rank}_{\mathscr{F}} Q = p, \tag{3.12}$$

where $\mathscr{F}$ is the field introduced above.

Proof. Necessity. Suppose that $\det Q = 0$ over $\mathscr{F}$. Then there exist $\alpha_1, \cdots, \alpha_p \in \mathscr{F}$, not all zero, such that $\sum \alpha_i * q_i = 0$ where q_i is the ith row of Q. Since each α_i is an element of the quotient field of $\mathscr{E}'(\mathbf{R}^-)/J$, this means that there exist $a_1, \cdots, a_p \in \mathscr{E}'(\mathbf{R}^-)/J$, not all zero, such that $\sum a_i * q_i = 0$. In view of the definition of the ideal J, we see that there exist $b_1, \cdots, b_p \in \mathscr{E}'(\mathbf{R}^-)$ such that $r(\sum b_i * q_i) < 0$ and $r(b_i) = 0$ for some i. Let $r_0 := r(\sum b_i * q_i)$, and put

$$\psi := \delta_{-r_0} * [b_1, \cdots, b_p]^T.$$

Then $r(\psi) > 0$ but $r(\psi^T * Q) \leqq 0$, and hence by Theorem 3.9 this system is not spectrally complete.

Sufficiency. Conversely, suppose that rank $Q = p$ over $\mathscr{F}$ but there exists $\psi \in (\mathscr{E}'(\mathbf{R}))^p$ such that $r_0 := r(\psi) > 0$ but $r(\psi^T * Q) \leqq 0$. Put

$$\varphi := \delta_{-r_0} * \psi.$$

Then φ belongs to $(\mathscr{E}'(\mathbf{R}^-))^p$, $r(\varphi) = 0$, and $r(\varphi^T * Q) < 0$. But this clearly contradicts rank $Q = p$ over $\mathscr{F}$. □

The following corollaries are now easy to prove.

COROLLARY 3.13. *Suppose that Q can be written in the form*

$$Q = Q_0 + Q_1, \tag{3.14}$$

where Q_0 is atomic at the origin and supp Q_1 *is contained in $(-\infty, t_0]$ for some $t_0 < 0$. Then Σ^Q is spectrally complete if and only if*

$$\operatorname{rank} \hat{Q}_0(\lambda) = p \tag{3.15}$$

for some $\lambda \in \mathbf{C}$.

Proof. In view of the well-known result on distributions atomic at the origin, each entry of Q_0 is a polynomial in the derivative δ' of the Dirac distribution [23]. Hence its Laplace transform is a polynomial in s. By Theorem 3.11, Σ^Q is spectrally complete if and only if rank $Q_0 = p$ over $\mathcal{F}$. In view of the definition of the ideal J, this is true if and only if det Q_0 is a nontrivial polynomial of δ'. This is clearly equivalent to condition (3.15). □

COROLLARY 3.16. *Consider the scalar case, that is, assume that the number of output channels $p = 1$. The system Σ^Q is spectrally complete if and only if $r(Q) = 0$.*

Proof. The proof is immediate from Theorem 3.11. □

Remark 3.17. The relationship of the above theorems in terms of the existing results is now clear. For simplicity, consider the scalar case. If the system Σ^Q is not spectrally complete, then $r(Q) < 0$; that is, Q can be written as $Q = \delta_{-a} * Q_1$ for some $a > 0$ and $Q_1 \in \mathcal{E}'(\mathbf{R}^-)$ with $r(Q_1) = 0$. We then easily see that $X^Q = L^2[0, a] \oplus X^{Q_1}$. Here $L^2[0, a]$ is nothing but the totality of "small solutions" considered in the literature [13], [20] (see Fig. 1).

Remark 3.18. The theorems above include some of the classical results on eigenfunction completeness. For example, consider the simplest case $\hat{Q}(s) := e^s - 1 (Q = \delta'_{-1} - \delta)$. It is easy to see that X^Q is the space of locally L^2 functions on $[0, \infty)$ of period 1. The set of eigenfunctions is $\{\exp(2n\pi jt);\ n = 0, \pm 1, \pm 2, \cdots\}$, and since $r(Q) = 0$, this set of eigenfunctions is complete in X^Q, as expected.

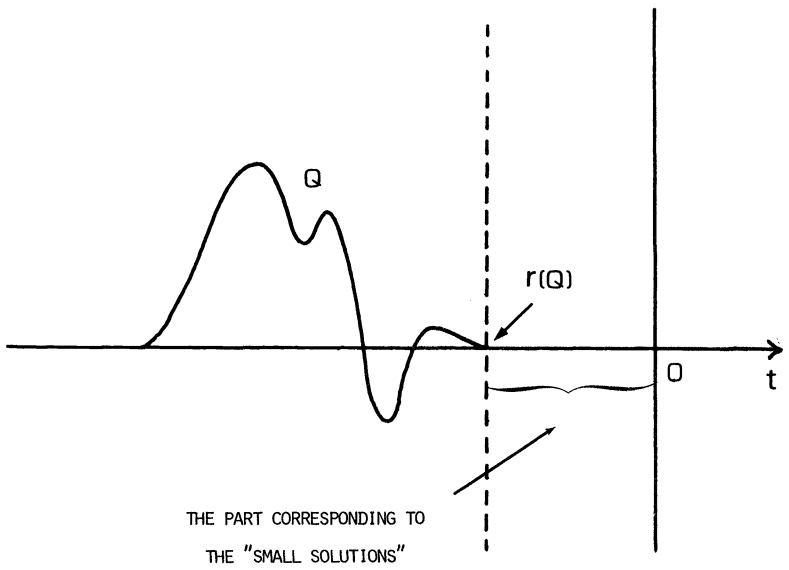

FIG. 1

4. Necessary and sufficient conditions for quasi-reachability. In this section, *we shall further assume that our impulse response matrix* $A = \pi(Q^{-1} * P)$ *is a regular distribution* (*i.e., locally integrable function*) *in a neighborhood of the origin.* Since all strictly causal impulse responses satisfy this requirement, this is not a very severe restriction. Then it follows that $\Psi := Q^{-1} * P - A$ must be a distribution belonging to $\mathscr{E}'(\mathbf{R}^-)$, and so is $Q * \Psi$. This implies that A can be rewritten as

$$A = Q^{-1} * P - \Psi = Q^{-1} * (P - Q * \Psi).$$

Obviously, the new pair $(Q, P - Q * \Psi)$ is also pseudo-rational, and by this modification, the truncation mapping π becomes unnecessary. In what follows, *we shall assume that* (Q, P) *satifies this condition*, i.e., $A = Q^{-1} * P$. (Note that the modification above does not change any of the conditions of the subsequent theorems.)

We are now ready to state and prove the following necessary and sufficient condition for quasi-reachability. As is the case with Theorem 3.11, there is no restriction on the type of delays involved.

THEOREM 4.1. *The system* Σ^Q *is quasi-reachable if and only if the following two conditions hold*:

(4.2) (i) $\operatorname{rank}[\hat{Q}(\lambda), \hat{P}(\lambda)] = p$ *for all* $\lambda \in \mathbf{C}$;

(4.3) (ii) $\operatorname{rank}_{\mathscr{F}}[Q, P] = p$.

Proof. Necessity. Condition (i) is clearly necessary, since it is a condition for spectral reachability (see [28]). Suppose that condition (ii) fails. Then there exist $a_1, \cdots, a_p \in \mathscr{F}$, not all zero, such that $a^T * [Q, P] = 0$ in $\mathscr{F}$ ($a = [a_1, \cdots, a_p]^T$). This readily implies that there exists $\psi \in (\mathscr{E}'(\mathbf{R}))^p$ such that $r(\psi) > 0$ but $r(\psi^T * Q) \leqq 0$, $r(\psi^T * P) \leqq 0$. If Σ^Q were quasi-reachable, there would exist sequences $\{R_n\}$ and $\{S_n\}$ of matrices over $\mathscr{E}'(\mathbf{R}^-)$ such that

$$Q * R_n + P * S_n \to \delta I_p, \tag{4.4}$$

i.e., Q and P are *approximately left coprime* by Theorem 2.13. Taking the convolution of (4.4) with ψ^T from the left, we see that the sequence $\psi^T * Q * R_n + \psi^T * P * S_n$ in $(\mathscr{E}'(\mathbf{R}^-))^p$ must converge to $\psi^T \notin (\mathscr{E}(\mathbf{R}^-))^p$ (because $r(\psi) > 0$). But this is clearly impossible.

Sufficiency. Conversely, suppose that the above two conditions hold. Since $\operatorname{rank}[Q, P] = p$ over $\mathscr{F}$, there exists a matrix K consisting only of zeros and ones such that $\operatorname{rank}[Q + PK] = p$ over $\mathscr{F}$. Furthermore,

$$\operatorname{rank}[\hat{Q}(\lambda) + \hat{P}(\lambda)K, \hat{P}(\lambda)] = \operatorname{rank}[\hat{Q}(\lambda), \hat{P}(\lambda)] = p \tag{4.5}$$

for all $\lambda \in \mathbf{C}$. According to Lemma 4.11 below, the pair $(Q + PK, P)$ is pseudo-rational. Thus we may consider the system Σ^{Q+PK} defined by the pair $(Q + PK, P)$; this system is quasi-reachable because it is spectrally reachable and, in addition to that, spectrally complete by Theorem 3.11. Therefore, by Theorem 2.13, there exist sequences $\{R_n\}$ and $\{S_n\}$ of matrices over $\mathscr{E}'(\mathbf{R}^-)$ such that

$$[Q + PK] * R_n + P * S_n \to \delta I_p. \tag{4.6}$$

Thus

$$Q * R_n + P * [KR_n + S_n] \to \delta I_p, \tag{4.7}$$

so that the pair (Q, P) is also approximately left coprime. Then again by Theorem 2.13, the system Σ^Q is quasi-reachable. □

In view of Theorem 2.13 we have simultaneously obtained the following theorem.

THEOREM 4.8. *The pair (Q, P) is approximately left coprime, i.e., there exist sequences of matrices R_n and S_n over $\mathscr{E}'(\mathbf{R}^-)$ of appropriate sizes such that*

$$Q * R_n + P * S_n \to \delta I \quad \text{in } (\mathscr{E}'(\mathbf{R}^-))^p$$

if and only if the above conditions (4.2) *and* (4.3) *hold.*

Proof. The proof is immediate from the fact that (Q, P) is approximately left coprime if and only if Σ^Q is quasi-reachable. □

The following corollaries are also direct consequences of Theorem 4.1.

COROLLARY 4.9. *Consider the scalar input/output system, i.e., the case $m = p = 1$. The system Σ^Q is quasi-reachable (and hence canonical) if and only if*

(i) $\operatorname{rank}[\hat{Q}(\lambda), \hat{P}(\lambda)] = 1$, *for all* $\lambda \in \mathbf{C}$;

(ii) $\max\{r(Q), r(P)\} = 0$.

Proof. The proof is immediate from Theorem 4.1. □

COROLLARY 4.10. *Consider a pair (Q, P) such that Q and P can be written in the form*

$$Q = Q_0 + Q_1, \qquad P = P_0 + P_1,$$

where Q_0 and P_0 are atomic at the origin, and supp Q_1, supp $P_1 \subset (-\infty, -t_0]$ *for some $t_0 > 0$. Then Σ^Q is quasi-reachable (and hence canonical) if and only if*

(i) $\operatorname{rank}[\hat{Q}(\lambda), \hat{P}(\lambda)] = p$ *for all* $\lambda \in \mathbf{C}$;

(ii) $\operatorname{rank}[\hat{Q}_0(\lambda), \hat{P}_0(\lambda)] = p$ *for some* $\lambda \in \mathbf{C}$.

Proof. Observe $Q \equiv Q_0$ and $P \equiv P_0$ modulo the ideal J. Then the result follows by the observation that $\operatorname{rank}[Q_0, P_0] = p$ over $\mathscr{F}$ if and only if the above condition (ii) holds as in the proof of Corollary 3.13. □

It remains only to prove that the pair $(Q + PK, P)$ considered above is pseudo-rational.

LEMMA 4.11. *Suppose that a pseudo-rational impulse response $A = Q^{-1} * P$ satisfies the hypothesis of the beginning of this section. Then the impulse response given by $[Q + PK]^{-1} * P$, where K is a constant matrix, is also pseudo-rational.*

Proof. We need to show the following:

(i) $Q + KP$ is invertible over $\mathscr{D}'_+$ with respect to convolution;

(ii) $\operatorname{ord}(\det(Q + KP)^{-1}) = -\operatorname{ord}(\det(Q + KP))$;

(iii) $[Q + PK]^{-1} * P$ is a valid impulse response.

Let us first show that $(\delta I + AK)^{-1}$ exists. Since we have assumed that A is a regular distribution (i.e., a locally integrable function) in a neighborhood of the origin, we may decompose A as $A = A_0 + A_1$, where A_0 is locally integrable and supp $A_1 \subset [t_0, \infty)$ for some $t_0 > 0$. It suffices to show that the Neumann series

$$\sum_{n=0}^{\infty} (-AK)^n \tag{4.12}$$

converges in $\mathscr{D}'_+$. In view of the topology of $\mathscr{D}'_+$ (Schwartz [23]), it is enough to see that the series (4.12) is convergent when applied to any C^∞-function φ with compact support. Note here that the support of A_1^n eventually becomes disjoint with that of φ. Then, expanding $(A_0 + A_1)^n$ by the binomial formula, and using the fact that the Neumann series (4.12) converges whenever A is a locally integrable function, we can easily show that (4.12) actually converges in $\mathscr{D}'_+$. Clearly, (4.12) gives $(\delta I + AK)^{-1}$ and

it is also a measure. Writing $[Q+PK]^{-1}$ as

$$[Q+PK]^{-1}=[\delta I+AK]^{-1}*Q^{-1}, \tag{4.13}$$

we see that $[Q+PK]^{-1}$ exists and is given indeed by (4.13).

Let us prove ord $(\det[Q+PK]^{-1})=-\text{ord}(\det[Q+PK])$. Since ord $(\alpha*\beta)\leqq$ ord $\alpha+$ord β is always valid, and since $\det[\delta I+AK]^{-1}$ is clearly a measure of order zero, we have

$$\text{ord}(\det[Q+PK]^{-1})\leqq\text{ord}(\det Q^{-1}), \tag{4.14}$$

by (4.13). Conversely, since $Q^{-1}=[\delta I+AK]*[Q+PK]^{-1}$ and since $\det[\delta I+AK]$ is also a measure of order zero, we have

$$\text{ord}(\det Q^{-1})\leqq\text{ord}(\det[Q+PK]^{-1}), \tag{4.15}$$

so that ord $(\det Q^{-1})=$ ord $(\det[Q+PK]^{-1})$. Now rewriting (4.13) as

$$Q+PK=Q*[\delta I+AK], \tag{4.16}$$

we also see that ord $(\det[Q+PK])=$ ord $(\det Q)$. Hence by the identity ord $(\det Q^{-1})=-$ord $(\det Q)$, we have ord $(\det[Q+PK]^{-1})=-$ord $(\det[Q+PK])$.

Finally, since $[\delta I+AK]^{-1}$ is a measure, $[Q+PK]^{-1}*P=[\delta I+AK]^{-1}*Q^{-1}*P=[\delta I+AK]^{-1}*A$ assumes the same regularity as A. Hence it is an impulse response matrix. □

5. Application to delay-differential systems. Consider the following neutral delay-differential system (with noncommensurable delays):

$$\frac{d}{dt}x(t)=F_0x(t)+\sum_{i=1}^{N}F_ix(t-h_i)+\sum_{i=1}^{N}F_{-i}\dot{x}(t-h_i)+Gu(t), \tag{5.1}$$

where $0<h_1<h_2<\cdots<h_N$. Let us temporarily take the output equation to be $y(t)=x(t-h_N)$. It is then appropriate to take Q and P as follows:

$$Q:=[\delta'_{-h_N}I-\delta_{-h_N}F_0-\textstyle\sum\delta_{-h_N+h_i}F_i-\sum\delta'_{-h_N+h_i}F_{-i}], \tag{5.2}$$

$$P:=G\delta. \tag{5.3}$$

Let $r_i:=h_i-h_{i-1}(h_0:=0)$, and $X:=\mathbf{R}^p\times\prod_{i=1}^{N}(L^2[0,r_i])^p$, where p is the dimension of x. Then, the realization Σ^Q corresponding to the factorization $Q^{-1}*P$ turns out to be of the following "M_2-type":

$$\begin{aligned}\frac{d}{dt}\begin{bmatrix}x\\z\end{bmatrix}&=\begin{bmatrix}F_0x+\sum_{i=1}^{N}[F_i+F_0F_{-i}]z_i(0)\\(\partial/\partial\theta)z(\theta)\end{bmatrix}+\begin{bmatrix}G\\0\end{bmatrix}u(t),\\&=F(x,z)^T+Gu(t),\end{aligned} \tag{5.4}$$

where $x\in\mathbf{R}^p$ and $z_i(\theta)\in(L^2[0,r_i])^p$. Here the domain $D(F)$ of the operator F is given by

$$\begin{aligned}D(F):=\{(x,z)\in X;\ z_i\in(W_2^1[0,r_i])^p,\ z_1(r_1)=x+\sum_i F_{-i}z_i(0),\\ z_k(r_k)=z_{k-1}(0),\ k=2,\cdots,N\}.\end{aligned} \tag{5.5}$$

(For details, see [30]; the difference between neutral and retarded is expressed in the definition of $D(F)$.)

THEOREM 5.6. *The neutral system* Σ^Q, *where* Q *and* P *are given by* (5.2) *and* (5.3), *is quasi-reachable if and only if*

(i) $\operatorname{rank}[\lambda e^{h_N\lambda} I - e^{h_N\lambda}F_0 - \sum e^{(h_N-h_i)\lambda}F_i - \sum \lambda e^{(h_N-h_i)\lambda}F_{-i},\ G] = p$ *for every* $\lambda \in \mathbf{C}$;

(ii) $\operatorname{rank}[F_N + \lambda F_{-N},\ G] = p$ *for some* $\lambda \in \mathbf{C}$.

Proof. The proof is immediate from Corollary 4.10. □

Theorem 5.6 extends the result of O'Connor and Tarn [17] to the noncommensurable delay case. It is also easy to consider the generalization to systems with distributed delays and to those with delays in input via Corollary 4.10, but the condition will look a little more involved for the distributed-delay case (in such a case, Theorem 4.1 is the final form). Note that there is no need to consider noncommensurable delays in any special way, since this is automatically taken care of by the framework itself.

Remark 5.7. According to Salamon [20, p. 39], the W_2^1-model of the neutral system (5.1) is nothing but the restriction of the M_2-realization (5.4) to the domain $D(F)$ of the infinitesimal generator F. In view of Proposition 2.22, this means that quasi-reachability is invariant for both systems. Thus the above condition also applies to the W_2^1-model.

The following example is taken from Salamon [20, Ex. 4.3.10].

Example 5.8. Consider the following neutral delay-differential system:

$$\dot{x}_1(t) = x_1(t) + \int_{-1}^{0} x_2(t+\tau)\,d\tau + \dot{x}_3(t-2),$$

$$\dot{x}_2(t) = x_2(t-1) + \int_{-1}^{0} x_3(t+\tau)\,d\tau,$$

$$\dot{x}_3(t) = \int_{-1}^{0} u(t+\tau)\,d\tau.$$

In conformity with the discussion there, take matrices $\hat{Q}$ and $\hat{P}$ as follows:

$$\hat{Q} := \begin{bmatrix} se^{2s} - e^{2s} & (e^{2s} - e^{s})/s & -s \\ 0 & se^{s} - 1 & (e^{s} - 1)/s \\ 0 & 0 & se^{s} \end{bmatrix}, \qquad \hat{P} := \begin{bmatrix} 0 \\ 0 \\ (e^{s} - 1)/s \end{bmatrix}. \tag{5.9}$$

It is easy to check that condition (i) of Theorem 4.1 is satisfied. Condition (ii) is also satisfied; i.e., the system is quasi-reachable. In fact, the matrices Q and P take the following form when considered over the ring $\mathscr{E}'(\mathbf{R}^{-1})/J$:

$$\tilde{Q} = \begin{bmatrix} 0 & 0 & -\delta' \\ 0 & -\delta & [1] \\ 0 & 0 & 0 \end{bmatrix}, \qquad \tilde{P} = \begin{bmatrix} 0 \\ 0 \\ [1] \end{bmatrix}, \tag{5.10}$$

where [1] denotes the equivalence class in $\mathscr{E}'(\mathbf{R}^{-1})/J$ of an element that is identically 1 for $-\varepsilon \leqq t \leqq 0$ and zero for $t \geqq 0$. Since this function has a jump of 1 at the origin, the determinant (with respect to convolution) formed by the second and third columns of $\tilde{Q}$ and $\tilde{P}$ is equal to $-\delta$, so that the pair $(\tilde{Q}, \tilde{P})$ assumes full rank over the quotient field $\mathscr{F}$. Hence by Theorem 4.1 the system Σ^Q is quasi-reachable.

Salamon [20] claims that this system is *not* quasi-reachable, contrary to the conclusion above. A detailed analysis shows that it is impossible to take a nontrivial small solution in his discussion [20, p. 155], and the system is in fact quasi-reachable. As can be seen from above, our test relying on (5.9) is much simpler than a discussion involving small solutions.

6. Concluding remarks. We have proved necessary and sufficient conditions for our standard observable realization Σ^Q to be quasi-reachable (and hence *canonical* in the sense of [27]). When we choose Q as a delay-differential operator, these results extend the known results on quasi-reachability of delay-differential systems. In particular, the results do not require *any restriction* on the type of delays (points or distributed). Aside from this, the approach here has the following advantages:

(1) It is suitable for the situation in which a priori association of a state space is not appropriate. A recently introduced control scheme called repetitive control is an example of this situation. The method here can be effectively used for proving the internal model principle for this nonclassical servo problem [8].

(2) The framework is closed under pole-zero cancellation. The existing state space approaches using M_2 or W_2^1, etc. do not possess this property.

(3) As a result of dealing with the input/output behavior, it clarifies the role of the distribution algebra $\mathscr{E}'(\mathbf{R}^-)$, which has not been considered in the literature. This is much more general than that considered by Manitius [14] (his algebra is not applicable to neutral systems).

(4) By virtue of the investigation of the convolution algebra structure, the obtained criterion (Theorem 4.1) is much simpler in the general case. The existing method requires an individual inspection of the nature of small solutions, which can often be quite complicated (see Example 5.8).

Appendix A.

Proof of Proposition 2.22. Sufficiency is obvious since $D(F)$ is always dense in X^Q (Yosida [31]).

Suppose that Σ^Q is quasi-reachable. Take any x in $D(F)$. Denote the reachable set of Σ_F^Q by X_R. We need to find a sequence in X_R converging to x with respect to the graph topology of $D(F)$. By hypothesis, the closure of X_R in X^Q is the whole space. Hence there exists a sequence $\{y_n\}$ in X_R such that $y_n \to Fx$ in X^Q. Let us again use g to denote the linear map sending inputs to states under the assumption of zero-initial state for Σ_F^Q. Let u_n be an input such that $y_n = g(u_n)$. Define

$$v_n(t) := \int_{-\infty}^{t} u_n(\tau)\, d\tau.$$

In view of the shift invariance of g and the fact that F is just the differential operator in X^Q, it is easy to verify that $g(v_n) \to x + x_0$ in X^Q, where x_0 is some constant function. Since x and $x + x_0$ belong to X^Q, x_0 does also. If x_0 were zero, there would be nothing left to prove, because we would have $Fg(v_n) = g((d/dt)v_n) = g(u_n) \to x$ and therefore find a desired sequence $\{g(v_n)\}$. Suppose $x_0 \neq 0$. This means that zero is an eigenvalue of F and x_0 is an eigenvector. By the assumption of quasi-reachability, every eigenvector must be reachable, and hence there exists v_0 such that $x_0 = g(v_0)$. Now let $w_n := v_n - v_0$. It readily follows that $g(w_n) \to x$, and $Fg(w_n) \to Fx$ in X^Q so that $g(w_n) \to x$ in $D(F)$. This completes the proof. □

Appendix B. We give a proof for Lemma 3.8. Let us first prove the following lemma.

LEMMA B1. *Let $\psi(s)$ and $q(s)$ be entire functions of exponential type, i.e., there exist C, $K > 0$ such that*

$$|\psi(s)| \leqq C e^{K|s|}, \qquad |q(s)| \leqq C e^{K|s|}.$$

Suppose that $\psi(s)q^{-1}(s)$ is an entire function. Then it is also an entire function of exponential type.

The result is nontrivial because, although $q(s)$ is of exponential type, there is no a priori guarantee that $q(s)^{-1}$ is of exponential type near a zero of $q(s)$. Indeed, the proof requires a deep result on the growth order of entire functions due to Lindelöf.

THEOREM B2 ([11], [1, Thm. 2.10.1]). *Let $f(s)$ be an entire function of order 1, i.e., for any $\varepsilon > 0$*

$$|f(s)| \leqq \exp(|s|^{1+\varepsilon})$$

for all sufficiently large $|s|$. Let $n(r)$ be the number of zeros of $f(s)$ in the circle $|s| \leqq r$, and let $\zeta_1, \cdots, \zeta_n, \cdots$ be the (nonzero) zeros of $f(s)$, counted according to multiplicities. Then $f(s)$ is of exponential type if and only if

(B3) (i) $n(r) \leqq O(r)$;
(ii) $S(r) := \sum_{|\zeta_n| \leqq r} 1/\zeta_n$

is bounded with respect to r.

Proof of Lemma B1. Let $f(s) := \psi(s) q(s)^{-1}$. Writing $\psi(s)$ and $q(s)$ in the form of the Hadamard factorization theorem [1, Thm. 2.7.1], we readily see that $f(s)$ is of order 1. Since all zeros of $q(s)$ must be cancelled by zeros of $\psi(s)$, it is clear that $n(r)$ of $f(s)$ is at most of $O(r)$. It remains to check the above condition (ii). Let $\{\xi_1, \cdots, \xi_n, \cdots\}$ and $\{\eta_1, \cdots, \eta_n, \cdots\}$ be the sets of (nonzero) zeros of $\psi(s)$ and $q(s)$, respectively. Let $S_1(r)$, $S_2(r)$ be the $S(r)$-functions for ψ and q, respectively. Again by the fact that all zeros of $q(s)$ are cancelled by those of $\psi(s)$, we have

$$\begin{aligned} |S(r)| &= \left| \sum_{|\xi_n| \leqq r} \frac{1}{\xi_n} - \sum_{|\eta_n| \leqq r} \frac{1}{\eta_n} \right| \\ &\leqq |S_1(r)| + |S_2(r)|, \end{aligned}$$

and the right-hand side is bounded by hypothesis on ψ, q. Hence $f(s)$ is also of exponential type. □

We are now ready to prove Lemma 3.8.

Proof of Lemma 3.8. Since $Q^{-1} = (\det Q)^{-1} * (\operatorname{adj} Q)$, we can apply the following argument to each entry of $\psi^T * Q^{-1}$, so we assume that ψ and Q are scalar distributions without loss of generality.

Case 1. Suppose that $f(s) = \psi(s) Q^{-1}(s)$ has only finitely many zeros. Then by the Hadamard factorization theorem [1, Thm. 2.7.1] we see that $f(s)$ must be of the form

$$f(s) = C \exp(as) \cdot \prod_{i=1}^{n} (s - \lambda_i), \tag{B4}$$

where C and a are suitable constants. Since $\psi * Q^{-1}$ is Laplace transformable, there exist b, $\beta \in \mathbf{R}$ such that its Laplace transform $e^{bs} f(s)$ is bounded by a polynomial in $|s|$ for $\operatorname{Re} s > \beta$ (Schwartz [22], [23]). In order that (B3) satisfy this requirement, the above constant a must be a real number. Then it is clear that its inverse Laplace transform is a distribution with compact support.

Case 2. Now consider the case in which $f(s)$ has infinitely many zeros. Since $\psi * Q^{-1}$ is Laplace transformable (i.e., $f(s)$ is inverse Laplace transformable), there exist b, $\beta \in \mathbf{R}$ such that $e^{bs} f(s)$ is bounded by a polynomial in $|s|$ for $\operatorname{Re} s > \beta$. We can take a vertical line $\operatorname{Re} s = \beta'$ such that there are infinitely many zeros of $f(s)$ except on this line. Collect suitably many of them, say $\lambda_1, \cdots, \lambda_n$, so that

$$\tilde{f}(s) := f(s)/(s - \lambda_1) \cdots (s - \lambda_n) \tag{B5}$$

remains an entire function and satisfies

$$|e^{bs} \tilde{f}(s)| \leqq C/|s|^2, \qquad \operatorname{Re} s = \beta'. \tag{B6}$$

By suitably shifting the coordinates, we may assume that this vertical line Re $s = \beta'$ is the imaginary axis. Denote this shifted function by $\bar{f}(s)$. It is easily seen that this shifted $\bar{f}(s)$ is square integrable on the imaginary axis and that it satisfies the estimate

$$|\bar{f}(s)| \leqq M \exp(K|s|) \tag{B7}$$

for some constants M and K. Then by a special version of the Paley-Wiener Theorem (Rudin [19, Thm. 19.3]), we see that $\bar{f}(s)$ is a Fourier transform of a function $F(t) \in L^2$ that has compact support in $[-K, K]$. This means that $\tilde{f}(s) = \bar{f}(s - \beta')$ is a Laplace transform of $F(t) \exp(\beta' t)$, which still has compact support. In view of (B4), $\psi * Q^{-1}$ can be obtained by applying the differential operator $(d/dt - \lambda_1) \cdots (d/dt - \lambda_n)$ to the inverse Laplace transform of $\tilde{f}(s)$, and hence $\psi * Q^{-1}$ has compact support. □

REFERENCES

[1] R. P. BOAS, JR., *Entire Functions*, Academic Press, New York, 1954.

[2] F. M. CALLIER AND C. A. DESOER, *An algebra of transfer functions for distributed linear time-invariant systems*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 651-662.

[3] C. A. DESOER AND M. VIDYASAGAR, Feedback Systems, Academic Press, New York, 1975.

[4] W. F. DONOGHUE, Distributions and Fourier Transforms, Academic Press, New York, 1969.

[5] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view*, J. Franklin Inst., 301 (1976), pp. 521-540.

[6] M. Q. JACOBS AND C. E. LANGENHOP, *Criteria for function space controllability of linear neutral systems*, SIAM J. Control Optim., 14 (1976), pp. 1009-1048.

[7] J. K. HALE, Theory of Functional Differential Equations, Springer-Verlag, Berlin, New York, 1977.

[8] S. HARA, Y. YAMAMOTO, T. OMATA, AND M. NAKANO, *Repetitive control system: a new type servo system for periodic exogenous signals*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 659-668.

[9] T. INOUE, M. NAKANO, T. KUBO, S. MATSUMOTO, AND H. BABA, *High accuracy control of a proton synchrotron magnet power supply*, Proc. Eighth Internat. Federation on Automatic Control, World Congress, XX: 216-221, 1981.

[10] E. W. KAMEN, P. P. KHARGONEKAR, AND A. TANNENBAUM, *Proper stable Bezout factorizations and feedback control of linear time-delay systems*, Internat. J. Control, 43 (1986), pp. 837-857.

[11] F. LINDELÖF, *Sur les fonctions entières d'ordre entier*, Ann. Sci. Ecole Norm. Sup., (3) 22 (1905), pp. 369-395.

[12] A. MANITIUS AND R. TRIGGIANI, *Function space controllability of linear retarded systems: a derivation from abstract operator conditions*, SIAM J. Control Optim., 16 (1978), pp. 599-645.

[13] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1-29.

[14] ———, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, SIAM J. Control Optim. 19 (1981), pp. 516-532.

[15] ———, *F-controllability and observability of linear retarded systems*, Appl. Math. Optim., 9 (1982), pp. 73-95.

[16] M. NAKANO AND S. HARA, *Microprocessor-based repetitive control*, Microprocessor-Based Control Systems, Reidel, Dordrecht, the Netherlands, 1986.

[17] D. A. O'CONNOR AND T. J. TARN, *On the function space controllability of linear neutral systems*, SIAM J. Control Optim., 21 (1983), pp. 306-329.

[18] H. R. RODAS AND C. E. LANGENHOP, *A sufficient condition for function space controllability of a linear neutral system*, SIAM J. Control Optim., 16 (1978), pp. 429-435.

[19] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[20] D. SALAMON, *Control and Observation of Neutral Systems*, Pitman, Boston, 1984.

[21] H. H. SCHAEFER, *Topological Vector Spacers*, Springer-Verlag, Berlin, New York, 1971.

[22] L. SCHWARTZ, *Méthodes mathématiques pour les sciences physiques*, Hermann, Paris, 1961.

[23] ———, *Théorie des distributions*, Deuxième édition, Hermann, Paris, 1966.

[24] F. TREVES, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York, 1967.

[25] M. VIDYASAGAR, H. SCHNEIDER, AND B. A. FRANCIS, *Algebraic and topological aspects of feedback stabilization*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 880-894.

[26] Y. YAMAMOTO, *Module structure of constant linear systems and its applications to controllability*, J. Math. Anal. Appl., 83 (1981), pp. 411-437.

[27] Y. YAMAMOTO, *Realization theory of infinite-dimensional linear systems, Parts* I *and* II, Math. Systems Theory, 15 (1981), pp. 55-77, pp. 169-190.

[28] ———, *Pseudorational input-output maps and their realizations: a fractional representation approach to infinite-dimensional systems*, SIAM J. Control Optim., 26 (1988), pp. 1415-1430.

[29] ———, *A note on linear input/output maps of bounded type*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 733-734.

[30] Y. YAMAMOTO AND S. UESHIMA, A new model for neutral delay-differential systems, Internat. J. Control, 43 (1985), pp. 465-472.

[31] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, New York, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 235-266, March 1989

A ROBUST ADAPTIVE MINIMUM VARIANCE CONTROLLER*

L. PRALY†, S.-F. LIN‡, AND P. R. KUMAR‡

Abstract. This paper addresses the twin questions of *performance* and *robustness* of an adaptive controller for single-input, single-output, linear, stochastic systems. The authors present an adaptive controller that has the following properties:

(1) Attaining *optimal* regulation and tracking in the *ideal*, minimum phase, known upper bound on system order, known sign and lower bound for the leading coefficient (b_0), positive real condition on noise case, and *self-tuning in a Cesaro sense* to a minimum variance regulator in the case of pure regulation.

(2) Providing *mean square stability* when the system is of minimum phase, with known upper bound on order *but not necessarily satisfying a positive real condition on the noise.*

(3) Providing *mean square stability* when the system is in a *graph topological neighborhood* (of computable size) of an ideal plant as in (1), and the statistical properties of the disturbance are violated.

(4) Continuing to *stabilize* the system when the *adaptation gain is prevented from vanishing.*

Key words. robustness, performance, adaptive control, optimal control, minimum variance control, graph topology, minimum variance regulator, self-tuning regulator

AMS(MOS) subject classification. 93C40

1. Introduction.

Over the past 15 years, stochastic adaptive control theory has seen much development. The notable pioneering contributions of Aström and Wittenmark [2] and Ljung [14], [15] analyzed, respectively, the *possible* equilibrium values of the parameters to which an adaptive control law could converge, and the *stability* properties of these equilibrium points. This set the stage for the subsequent rigorous development of the foundations of the asymptotic theory of the so-called self-tuning controllers.

In 1981, Goodwin, Ramadge, and Caines [6] were able to successfully use some extensions of the martingale convergence theorem to show the convergence of a certain stochastic Lyapunov function. They were thus able to establish that for a variety of stochastic gradient algorithms the time average of the squared tracking error is almost surely optimal, a property we shall refer to as *self-optimality.* These results were then extended by similar arguments to some other algorithms; for example, an adaptive controller based on a modified least-squares estimate was analyzed by Sin and Goodwin [24]. In 1985, Becker, Kumar, and Wei [3] addressed the issue of convergence of the parameter estimates, and, in so doing, they also established the convergence of the adaptive regulator. By exploiting some geometric properties of the parameter estimate sequence, and some subsequent probabilistic analysis, they were able to show that while the parameter estimates converge almost surely (a.s.), they do *not* converge to their true values. Instead, the parameter estimate vector converges to a *random scalar multiple* of the true parameter vector. However, since the particular control law used for the regulation problem employs only *ratios* of estimates of individual parameters,

* Received by the editors July 20, 1987; accepted for publication (in revised form) April 18, 1988.

† CAI/ENSMP, 77305 Fontainebleu Cedex, France. This author is a member of the Groupe de Recherche Coordonée-Systèmes Adaptatifs en Robotique, Traitement du Signal et Automatique of the CNRS.

‡ Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois, 1101 W. Springfield Avenue, Urbana, Illinois 61801. The research of this author was supported in part by National Science Foundation grant ECS-84-14676, in part by the U.S. Army Research Office under contract DAAL03-88-K-0046, and in part by the Joint Services Electronics Program under contract N00014-84-C-0149.

the adaptive control law remains invariant under scaling of the parameter estimates. Hence convergence of the adaptive regulator to the true optimal regulator takes place almost surely. This result therefore proved the so-called *self-tuning* property of the adaptive regulator. Recently, this self-tuning property has been extended by Kumar and Praly [11] to the *tracking* problem, where the goal is not to regulate the system output to stay close to zero, but to *track* a given *reference trajectory* while *optimally rejecting* the noise entering into the system. The essential difference between the regulation and tracking problems is that in the former problem it is not necessary to estimate the coefficients of the colored noise polynomial (in the ARMAX representation of the system), while in the latter it is necessary to do so if we want to track arbitrary reference trajectories. Since an additional number of parameters have to be estimated in the tracking problem, it turns out that more "excitation" of the system is needed. This excitation is in turn guaranteed if the reference trajectory is sufficiently exciting of appropriately high order (see Kumar and Praly [11]). In many practically important situations, such as, for example, the *set-point problem*, however, the reference trajectory may only be a *constant* level, which is sufficiently exciting of *order one*, or some other trajectory that has a low order of excitation. In these situations, it turns out that not all the coefficients of the colored noise polynomial need be explicitly estimated, but rather knowledge of a smaller set of parameters derived from the coefficients is adequate. This allows the design of an adaptive tracker that uses a smaller dimension parameter estimator (which is still of larger dimension than is needed in the regulation problem). Such smaller dimensional adaptive trackers have also been proved to be *self-tuning* (see Kumar and Praly [11]).

The successful results quoted above essentially show that the adaptive regulators and trackers tune themselves to *optimal* regulators and trackers for the unknown system. Inevitably, such results are crucially dependent on making some "exact" assumptions about the unknown system being controlled. In particular, for such exact asymptotic optimality and strong convergence results to hold, it has been assumed that the stochastic system being controlled is linear, of minimum phase, of *known order*, and the disturbance entering into the system is a *stochastic process* satisfying some specified *statistical* properties.

Assumptions of the above type have been called "ideal" assumptions, and questions have been raised, especially in *deterministic* adaptive control (see Egardt [5], Rohrs, Valavani, Athans, and Stein [22], and Ioannou and Kokotovic [7]), about whether the adaptive controllers designed on the basis of these assumptions, and for which a successful "ideal" theory has been built, are *robust* with respect to these assumptions. Specifically, do "small' violations of these assumptions lead to drastically different behavior from that predicted by the ideal theory?

The *order* requirement arises since we must choose the dimension of the adaptive regulator before we can tune it. However, the true system need not necessarily be (and is frequently not) of the exact order that is assumed. It is well known that "small perturbations" of an nth-order linear system can lead to systems of arbitrarily high order.

Regarding the disturbance, the assumption made is that it is a *stochastic process* with a *rational spectral density*, and thus is representable as the output of a system driven by *white noise*. Moreover, the *order* of this "coloring" filter is assumed known. Finally, it is also assumed that the noise satisfies a certain "positive real" condition. This is essentially a requirement that the disturbance be close to a white noise and be not too colored. However, none of these assumptions need be strictly satisfied in practice. The positive realness condition also arises in recursive system identification using the pseudolinear regression method (see Solo [25], Ljung and Söderström [16],

and Kumar and Varaiya [12]). It is basically a *pseudogradient condition* (see Ljung and Söderström [16] and Kumar [10]) guaranteeing that the direction in which the parameter estimates are recursively adjusted (in the types of recursive identification algorithms being employed) is appropriate.

Regarding the minimum phase restriction, it is well known (see Aström [1], Peterka [17], Shaked and Kumar [23], and Kumar and Varaiya [12]) that when a *stationary* control law that minimizes the output variance is used to control the system, then the control actions used become unbounded if the system is of *nonminimum phase.* However, for *adaptive control* where a *nonstationary,* nonlinear control law is used, it is not necessary that the minimum phase assumption be satisfied in order for the control inputs to be bounded. Hence the minimum phase assumption is a restrictive condition; it is easily violated by a very fast unstable zero that corresponds to a very small numerator perturbation of the transfer function.

Much attention has therefore been given in recent years to the issue of *robust adaptive control,* especially in deterministic adaptive control, to determine under what conditions signals in the system remain bounded under violations of assumptions (for example, see [8], [9], [20]). In the adaptive control of *stochastic* systems, however, noise is an essential feature of the system, and it is of interest not only to guarantee boundedness of signals, but it is also important to reject the noise optimally, or at least much of it. Thus, *performance* of the adaptive control algorithm in rejecting the corrupting noise, and thus tracking the desired reference trajectory with *small* tracking error, is also an important goal in stochastic adaptive control.

In this paper, therefore, we address the twin questions of *performance as well as robustness* of adaptive control laws for linear stochastic systems. In particular, we address the issue of adaptive controllers that are *performance-optimal* when the ideal assumptions are satisfied, *and* that are *robust with respect to perturbations* of the system from the ideal assumptions.

We will consider two types of *perturbations* of the system from optimality. First we consider perturbations of the coefficients of the colored noise polynomial that can be large and that allow gross violation of the positive real assumption. This problem has been treated by Egardt [5] for bounded noise and extended in Praly [18] for mean-square bounded noise.

Second, we consider *system perturbations.* Vidyasagar [26] has identified the appropriate topology on the set of linear systems, called the *graph topology,* which is the *weakest topology* such that there is a stabilizing linear controller for a nominal ideal system that remains stabilizing, and such that the closed-loop transfer function is continuous (uniformly over all frequencies) when perturbations with respect to this topology are allowed. Thus, for any given weaker topology which thus allows *more* perturbations, there is not necessarily any single linear control law that continues to maintain stability. Since self-tuning or adaptive control is really an online or real-time search over the space of linear controllers, we cannot expect to do better than allow for perturbations with respect to this graph topology. Thus while (nonadaptive) linear controllers are designed for perturbations with respect to the graph topology from a *given nominal system,* adaptive control laws should be designed to maintain stability with respect to the graph topology from all possible nominal systems. This indeed is the goal of this paper. We will achieve it by extending the approach of Praly [19] to the vanishing gain case.

Last, asymptotic optimality and convergence results for adaptive controllers rely on adaptive parameter adjustment schemes that use an asymptotically vanishing step-size, i.e., the gain converges to zero. However, to maintain the ability to adapt, the

gain should be nonvanishing. Thus we also need to analyze the effect of nonvanishing gain on the ideal adaptive control algorithm.

In this paper we therefore exhibit an adaptive controller for linear stochastic systems that is *optimal for all ideal plants*, and remains stable with respect to violations of the positive real condition, and with respect to perturbations of the system, in the graph topology, from all ideal plants. Moreover, we show that stability is preserved when the gain is prevented from going to zero.

Specifically, we present an adaptive controller for which we prove the following performance and robustness properties:

(1) Attaining *optimal* regulation and tracking in the *ideal* case when the system is of minimum phase with a known upper bound on the system order, and when the coefficients of the colored noise polynomial satisfy a positive real condition (Theorem 5.1). In the case of the *regulation* problem, we also show that the adaptive controller *self-tunes in a Cesaro sense* to minimum variance regulator (Theorem 5.2).

(2) Providing *mean-square stability* when the system is of minimum phase with a known upper bound on the system order but *does not necessarily satisfy a positive real condition* (Theorem 4.6).

(3) Providing *mean-square stability* when the system is in a *graph topological neighborhood of computable size* of an ideal system as in (1) (Theorem 6.8).

(4) *Continuing to stabilize* the system when the *adaptive gain is prevented from vanishing to zero* (Theorem 7.7).

There are still many unresolved questions. Maybe the most important is to determine whether adaptive controllers without the modifications we have used are already robust, even though our modifications are well motivated. Moreover, we have not really been able to deal with the removal of the minimum phase assumption, even though, as we will show later, our adaptive controller is robust with respect to graph topological perturbations that do result in nonminimum phase systems.

2. The adaptive controller. In this section we present our adaptive controller. In the next five sections we analyze the effect of the adaptive controller when it is applied to a variety of systems satisfying varying assumptions. (Thus we are reversing the usual order of presentation, where the intended systems are first described before the adaptive controllers are defined!)

We will suppose that the system under control is a single-input, single-output system with input sequence $u(t)$ and output sequence $y(t)$. We will also suppose the following:

(A2.i) The reference trajectory $y^m(t)$ is bounded.

There are several fixed parameters that are chosen a priori. We choose the following:

(A2.ii) Three integers n_R, n_S, and n_C (which describe the dimensions of our adaptive controller, but not necessarily those of the system);

(A2.iii) Two positive numbers $0 < \lambda_0 < \lambda_1$ (which serve as bounds on certain eigenvalues);

(A2.iv) Three positive numbers $\rho > 0$, $\sigma_0 > 0$, and $K > 0$;

(A2.v) A parameter vector θ^c of dimension $(n_R + n_S + n_C + 2)$ whose first component is larger than or equal to σ_0;

(A2.vi) An integer $d \geqq 1$ (which models the delay but may not be equal to it).

We use the *regression vector* $\phi(t)$ defined as

$$\phi(t) := (u(t), \cdots, u(t-n_S), y(t), \cdots, y(t-n_R), y^m(t+d-1), \cdots, y^m(t+d-n_C))^T.$$

Given $\theta(n)$, $F(n)$, and $\rho(n) > 0$ for all $n \leqq t-1$, and having applied a new control input $u(t-1)$ and observed a new output $y(t)$, we *recursively* define the adaptive controller as follows:

$$\rho(t) := \rho(t-1) + \max(\rho, \|\phi(t-d)\|^2), \quad t \geqq 1 \quad \text{(we choose } \rho(t) = 0 \text{ for } t \leqq 0), \tag{2.1}$$

$$\bar{\phi}(t-d) := \frac{\phi(t-d)}{\rho^{1/2}(t)}, \tag{2.2}$$

$$g(t) := \frac{1}{1 + \bar{\phi}^T(t-d)F(t-d)\bar{\phi}(t-d)}, \tag{2.3}$$

$$e(t) := y(t) - \theta^T(t-d)\phi(t-d), \tag{2.4}$$

$$\bar{e}(t) := \frac{e(t)}{\rho^{1/2}(t)}, \tag{2.5}$$

$$F^1(t) := F(t-d) - g(t)F(t-d)\bar{\phi}(t-d)\bar{\phi}^T(t-d)F(t-d), \tag{2.6}$$

$$F(t) := \left(1 - \frac{\lambda_0}{\lambda_1}\right)F^1(t) + \lambda_0 I \quad \text{(we choose } \lambda_0 I \leqq F(0) \leqq \lambda_1 I), \tag{2.7}$$

$$\theta^1(t) := \theta(t-d) + g(t)F(t-d)\bar{\phi}(t-d)\bar{e}(t), \tag{2.8}$$

$$\theta^2(t) := \theta^1(t) + \max(0, \sigma_0 - s_0^1(t)) \frac{F_1(t)}{F_{11}(t)} \tag{2.9}$$

where

$$s_0^1(t) := \text{first element of the vector } \theta^2(t),$$

$$F_1(t) := \text{first column of the matrix } F(t),$$

$$F_{11}(t) := (1,1)\text{th element of } F(t),$$

$$\theta(t) := \theta^c + (\theta^2(t) - \theta^c) \min\left(1, \frac{K\lambda_1}{\lambda_0 \|\theta^2(t) - \theta^c\|}\right). \tag{2.10}$$

Finally, the control input is defined implicitly through

$$\theta^T(t)\phi(t) = y^m(t+d). \tag{2.11}$$

Explanation of adaptive control algorithm. There are essentially only three features of our adaptive control law that are different from the usual adaptive control laws.

Normalization. The sequence $\rho(t)$ is a *normalization* (or *scaling*) sequence. The vector $\bar{\phi}(t-d)$ obtained by normalizing (i.e., dividing) $\phi(t-d)$ by $\rho^{1/2}(t)$ is then the *normalized* regression vector, and similarly $\bar{e}(t)$ is the *normalized prediction error.* These normalized signals are then used to update the parameter estimates.

Condition number bounding. The matrix $F(t)$ is what is usually called the "covariance matrix." It is well known in recursive identification (see Lai and Wei [13] and Kumar and Varaiya [12]) that if the condition number of the so-called "covariance matrix" remains bounded as $t \to \infty$, then the parameter estimates converge to their true values. Equation (2.7) ensures that the eigenvalues of $F(t)$ remain within the interval $[\lambda_0, \lambda_1]$, thereby keeping the condition number uniformly bounded. (In fact, as the reader can verify, any $F(t) \geqq F^1(t)$ satisfying the property that its eigenvalues lie in the interval $[\lambda_0, \lambda_1]$ can be used.)

Parameter estimate projection. Finally there is a set of two modifications that ensure that the parameter estimates are kept bounded, while at the same time making sure that the first component of the vector $\theta(t)$ (which is an estimate of the so-called "high-frequency gain" of the system) is kept positive and bounded below. This is done in two stages. The first stage, (2.9), ensures that the first component is larger than σ_0. The second stage, (2.10), keeps the parameter estimates inside the sphere with center θ^c and radius $K\lambda_1/\lambda_0$ by projecting them radially onto the surface of the sphere whenever they wander outside.

Remarks on modifications. The reasonableness of the modifications of normalization and eigenvalue bounding can be seen from the following calculation. Normal *unmodified* adaptive control laws using least-squares parameter estimates would use the (d interlaced) recursions

$$\theta(t)=\theta(t-d)+\frac{R^{-1}(t-d)\phi(t-d)}{1+\phi^T(t-d)R^{-1}(t-d)\phi(t-d)}(y(t)-\theta^T(t-d)\phi(t-d)),$$

$$R(t)=R(t-d)+\phi(t-d)\phi^T(t-d).$$

These recursions are clearly equivalent to

$$\theta(t)=\theta(t-d)+\frac{\left(\frac{R(t-d)}{\rho(t)}\right)^{-1}\frac{\phi(t-d)}{\rho^{1/2}(t)}}{1+\frac{\phi^T(t-d)}{\rho^{1/2}(t)}\left(\frac{R(t-d)}{\rho(t)}\right)^{-1}\frac{\phi(t-d)}{\rho^{1/2}(t)}}\frac{(y(t)-\theta^T(t-d)\phi(t-d)}{\rho^{1/2}(t)},$$

i.e.,

$$\theta(t)=\theta(t-d)+\frac{\left(\frac{R(t-d)}{\rho(t)}\right)^{-1}\bar{\phi}(t-d)}{1+\bar{\phi}^T(t-d)\left(\frac{R(t-d)}{\rho(t)}\right)^{-1}\bar{\phi}(t-d)}\bar{e}(t).$$

Thus we see that modified adaptive control uses $F(t-d)$ instead of $(R(t-d)/\rho(t))^{-1}$. This is reasonable since $R(t-d)/\rho(t)\leqq I$, and $F(t-d)$ also has a lower-bounded minimum eigenvalue. Hence both $R^{-1}(t-d)/\rho(t)$ and $F(t-d)$ are of the *same order* and grow at the same rate. Last, the bounding of the *maximum eigenvalue* of $F(t-d)$ is a reasonable effort at keeping the condition number bounded.

An intuitive rationale for the introduction of normalization is the following. Let us consider the case where the system is not of the order assumed. Then, generally we can assume that the system can be represented in the following form (which also allows infinite-dimensional systems):

$$y(t)=ay(t-1)+bu(t-1)+\sum_{i=2}^{t}(\alpha_i y(t-i)+\beta_i u(t-i))$$

where the summation represents the portion of the system dynamics that has not been modeled. Then, under the assumption that $n_S=n_R=0$, we have $\phi(t-1)=(u(t-1),y(t-1))$, and so for any $\theta=(\theta_1,\theta_2)^T$,

$$y(t)-\phi^T(t-1)\theta=(a-\theta_2)y(t-1)+(b-\theta_1)u(t-1)+\sum_{i=2}^{t}(\alpha_i y(t-i)+\beta_i u(t-i)).$$

This *modeling error* may be unbounded irrespective of the choice of θ. However, the neglected component can be bounded by

$$\left|\sum_{i=2}^{t}(\alpha_i y(t-i)+\beta_i u(t-i))\right| \leqq \sqrt{2}\left(\sum_{i=2}^{t}\alpha_i^2+\beta_i^2\right)^{1/2}\left(\sum_{i=2}^{t} y^2(t-i)+u^2(t-i)\right)^{1/2}$$

by the Cauchy–Schwarz inequality. Noting that $\sum_{i=2}^{t}(y^2(t-i)+u^2(t-i)) \leqq \rho(t)$, where $\rho^{1/2}(t)$ is the *normalization factor*, we see that $|\bar{y}(t)-\bar{\phi}^T(t-1)\theta| \leqq M_0$, when $\{\alpha_i\}$ and $\{\beta_i\}$ are in l_2. Hence the error due to mismodeling is bounded when we use the normalized quantities instead of the original variables. This is the heuristic reason for our use of normalization.

The purposeful bounding of the parameter estimates (by keeping them in a certain sphere), which is our last modification, does not cause any problems, at least when the "true parameter vector" is known to satisfy a similar bound, thus allowing convergence of the parameter estimates to their "true values" if that is necessary. As we show later, there need not even be a "true parameter vector" for this modification to be reasonable. In fact, Egardt [5] has shown that some sort of parameter boundedness is necessary for good behavior. Similarly, keeping the first component of the parameter estimates bounded below is tolerable at least when the true parameter vector also has the same lower bound on its first component.

It should be noted that our bounding of the eigenvalues of $F(t)$ is somewhat similar to the case of the *stochastic gradient algorithm* (see Becker, Kumar, and Wei [3]). In fact, the stochastic gradient algorithm is a special case of our modified adaptive controller that is obtained when we choose $\lambda_0=\lambda_1$ in (2.7). In general, however, we expect that the initial transient performance of the adaptive controller will be closer to the least-squares algorithm, but that the asymptotic convergence rate will be governed by that of the gradient algorithm, although we have not been able to establish either of these results analytically.

The modifications present in our adaptive controller, which were first proposed in Praly [21], therefore, all stem from reasonable motivations. In what follows we actually show the power of these modifications in a variety of situations.

3. Some properties of the adaptive controller. Interestingly enough (and very useful to us), the adaptive controller that we defined earlier satisfies some useful conditions *irrespective* of the system to which it is applied.

Let us define Θ as the intersection of the closed sphere with center θ^c and radius K, with the closed half-space $s_0 \geqq \sigma_0$ (where s_0 = first component of vector $\theta \in \Theta$). Note that by construction (see (A2.v)) θ^c belongs to Θ. For any $\theta \in \Theta$, we define the *prediction error* by

$$w_\theta(t) := y(t)-\theta^T\phi(t-d) \tag{3.1}$$

and its *normalized* version by $\bar{w}_\theta(t) := w_\theta(t)/\rho^{1/2}(t)$.

We wish to emphasize that the results of this section are obtained without any assumptions on $\bar{w}_\theta(t)$. The following preliminary results are of much interest, and will be very useful to us. Since they are a direct consequence of our definitions, their proofs are omitted.

LEMMA 3.1.

(i) $1 \geqq g(t) \geqq 1/(1+\lambda_1)$;

(ii) *If* $\theta \in \Theta$, *then* $\|\theta(t)-\theta\| \leqq K_1$, *for some constant* K_1;

(iii) $\rho(T) \geqq \sum_{t=1}^{T}\|\phi(t-d)\|^2$.

Proof. The proof is trivial. □

It should be noted that $g(t)$ is the only eigenvalue of the matrix $[I - g(t)F(t-d)\bar{\phi}(t-d)\bar{\phi}^T(t-d)]$ that is not equal to 1. Since (2.8) can be rewritten as $\theta^1(t) = [I - g(t)F(t-d)\bar{\phi}(t-d)\bar{\phi}^T(t-d)]\theta(t-d) + g(t)F(t-d)\bar{\phi}(t-d)\bar{y}(t)$, it is then clear that $g(t)$ tells us how *contractive* the homogeneous part of this update equation is, and (i) provides a lower bound on the rate of convergence of the parameters. Statement (ii) above merely makes note of the fact that $\theta(t)$ is kept bounded.

LEMMA 3.2. *Define a Lyapunov function* $V_\theta(t) := (\theta(t)-\theta)^T F^{-1}(t)(\theta(t)-\theta)$, *for* $\theta \in \Theta$. *Then*

$$V_\theta(t) \leqq V_\theta(t-d) + \bar{w}_\theta^2(t) - g(t)\bar{e}^2(t).$$

Proof.

Step 1. $(F^1(t))^{-1} = F^{-1}(t-d) + \bar{\phi}(t-d)\bar{\phi}^T(t-d)$. After some algebra and (3.1), we have

$$(\theta^1(t)-\theta)^T(F^1(t))^{-1}(\theta^1(t)-\theta) = V_\theta(t-d) + \bar{w}_\theta^2(t) - g(t)\bar{e}^2(t). \tag{3.2}$$

Since $(F^1(t))^{-1} \geqq F^{-1}(t)$ and because of (3.2), we have

$$(\theta^1(t)-\theta)^T F^{-1}(t)(\theta^1(t)-\theta) \leqq V_\theta(t-d) + \bar{w}_\theta^2(t) - g(t)\bar{e}^2(t). \tag{3.3}$$

Step 2. Let $\Delta'(t) = (\theta^2(t)-\theta)^T F^{-1}(t)(\theta^2(t)-\theta) - (\theta^1(t)-\theta)^T F^{-1}(t)(\theta^1(t)-\theta)$. Then some algebra yields $\Delta'(t) = (\theta^2(t) + \theta^1(t) - 2\theta)^T F^{-1}(t)(\theta^2(t) - \theta^1(t))$.

Now we consider two cases.

Case 1. If $\sigma_0 \leqq s_0^1(t)$ then $\theta^1(t) = \theta^2(t)$ and so $\Delta'(t) = 0$.

Case 2. If $\sigma_0 > s_0^1(t)$ then

$$\Delta'(t) = (\theta^2(t) + \theta^1(t) - 2\theta)^T e_1 \frac{\sigma_0 - s_0^1(t)}{F_{11}(t)}$$

$$= \frac{\sigma_0 - s_0^1(t)}{F_{11}(t)}[s_0^1(t) + (\sigma_0 - s_0^1(t)) + s_0^1(t) - 2s_0] \leqq 0 \quad (\text{since } s_0^1(t) < \sigma_0 < s_0)$$

where $e_1 = (1, 0, \cdots, 0)^T$.

Hence, in any case we have

$$(\theta^2(t)-\theta)^T F^{-1}(t)(\theta^2(t)-\theta) \leqq V_\theta(t-d) + \bar{w}_\theta^2(t) - g(t)\bar{e}^2(t). \tag{3.4}$$

Step 3. For convenience, let M_1 and d_1 denote

$$M_1 := (\theta^2(t)-\theta)^T F^{-1}(t)(\theta^2(t)-\theta) \quad \text{and} \quad d_1 := K\frac{\lambda_1}{\lambda_0}\frac{1}{\|\theta^2(t)-\theta^c\|}.$$

Now consider two cases again.

Case 1. If $d_1 \geqq 1$, then $\theta(t) = \theta^2(t)$ and so $M_1 = V_\theta(t)$.

Case 2. If $d_1 < 1$, using (2.10) and the Cauchy-Schwarz inequality, then

$$M_1 - V_\theta(t) = (\theta^2(t) + \theta(t) - 2\theta)^T F^{-1}(t)(\theta^2(t) - \theta(t))$$

$$\geqq (1-d_1^2)\|\theta^2(t)-\theta^c\|^2\frac{1}{\lambda_1} - 2\frac{K}{\lambda_0}(1-d_1)\|\theta^2(t)-\theta^c\|$$

$$= \frac{\|\theta^2(t)-\theta^c\|^2}{\lambda_1}[1 - d_1^2 - 2(1-d_1)d_1] \geqq 0.$$

Hence, in any case, $M_1 \geqq V_\theta(t)$ and the result follows. □

The above recursive bound on the "Lyapunov function" will be useful subsequently.

LEMMA 3.3.

(i) $s_0(t) \geqq \sigma_0$;

(ii) $\|\theta(t)\| \leqq \|\theta^c\| + K(\lambda_1/\lambda_0) =: R$;

(iii) $e(t) = y(t) - y^m(t)$;

(iv) $\|\theta(t) - \theta(t-d)\| \leqq \sqrt{\lambda_1}(1+\sqrt{\lambda_1}/\sqrt{\lambda_0})|\bar{e}(t)|$;

(v) *For any* $\theta \in \Theta$, $e^2(t) \leqq \rho(t)(1+\lambda_1)(V_\theta(t-d) - V_\theta(t)) + (1+\lambda_1)w_\theta^2(t)$ *and* $0 \leqq V_\theta(t) \leqq V_4 := 1/\lambda_0(K + K(\lambda_1/\lambda_0))^2$.

Proof. Formulas (i)-(iii) follow almost by definition.

(iv) Because $\|g(t)F(t-d)\bar{\phi}(t-d)\| \leqq \sqrt{\lambda_1}/2$, it follows that

$$\|\theta^1(t) - \theta(t-d)\| \leqq \frac{\sqrt{\lambda_1}}{2}|\bar{e}(t)|. \tag{3.5}$$

From the algorithm, we can easily see that

$$(\theta^2(t) - \theta^1(t))^T F^{-1}(t)(\theta^2(t) - \theta^1(t)) \leqq \frac{(\sigma_0 - s_0^1(t))^2}{\lambda_0}, \tag{3.6}$$

$$\|\theta^2(t) - \theta^1(t)\|^2 \leqq \frac{\lambda_1}{\lambda_0}\|\theta(t-d) - \theta^1(t)\|^2, \tag{3.7}$$

$$\|\theta(t) - \theta^2(t)\| \leqq \|\theta^2(t) - \theta(t-d)\|. \tag{3.8}$$

Using (3.7) and (3.5), we have

$$\|\theta^2(t) - \theta^1(t)\|^2 \leqq \frac{\lambda_1^2}{4\lambda_0}|\bar{e}(t)|^2. \tag{3.9}$$

Combining (3.8), (3.9), and (3.5), we have

$$\|\theta(t) - \theta(t-d)\| \leqq \|\theta(t) - \theta^2(t)\| + \|\theta^2(t) - \theta(t-d)\| \leqq 2\|\theta^2(t) - \theta(t-d)\|$$

$$\leqq \sqrt{\lambda_1}\left(1 + \frac{\sqrt{\lambda_1}}{\sqrt{\lambda_0}}\right)|\bar{e}(t)|.$$

(v) From Lemmas 3.2 and 3.1(i),

$$\frac{\bar{e}^2(t)}{1+\lambda_1} \leqq g(t)\bar{e}^2(t) \leqq V_\theta(t-d) - V_\theta(t) + \bar{w}_\theta^2(t),$$

and the bound on $e^2(t)$ follows readily. On the other hand,

$$V_\theta(t) \leqq \frac{1}{\lambda_0}\|\theta(t) - \theta\|^2 \leqq \frac{1}{\lambda_0}\left(\|\theta^c - \theta\| + K\frac{\lambda_1}{\lambda_0}\right)^2$$

and the claimed bound follows, since $\|\theta^c - \theta\| \leqq K$ due to the requirement that $\Theta \in \Theta$. □

The first result above merely states that the subsequent projection onto the surface of the sphere continues to preserve the property (i). The fourth result above gives a bound on the *increments* of the parameter estimates in terms of the normalized errors, while the last result gives a *bound* on the normalized errors themselves.

This last result is fundamental. It shows that insofar as the norms of the sequences are concerned, the adaptation law may be regarded as a *static gain* operator with inputs $w_\theta(t)$, $\sqrt{\rho(t)}$ and output $e(t)$. The gain from $w_\theta^2(t)$ to $e^2(t)$ is simply $(1+\lambda_1)$, which increases as the *speed* of adaptation measured by the largest eigenvalue λ_1 is

concerned. It tells us that the error given by the parameter estimates will be smaller than $\sqrt{1+\lambda_1}$ times the error given by *any* vector $\theta \in \Theta$. The gain from $\rho(t)$ to $e^2(t)$ is $(1+\lambda_1)(V_\theta(t-d)-V_\theta(t))$. Suppose now that, due to the boundedness of $V_\theta(\cdot)$, the "mean" value of $V_\theta(t-d)-V_\theta(t)$ is close to zero. Then boundedness of $e^2(t)$ will follow from the small gain theorem [4] if the operator $e(t) \to \sqrt{\rho(t)}$ has bounded gain and the operator $e(t) \to w_\theta(t)$ is an operator whose gain multiplied by $\sqrt{1+\lambda_1}$ is smaller than 1. Moreover, since this result holds for *all* $\theta \in \Theta$, we have

$$\frac{1}{T}\sum_{i=t}^{t+T} e^2(i) \leqq \frac{1}{T}\sum_{i=t}^{t+T} \rho(i)(1-\lambda_1)(V_\theta(i-d)-V_\theta(i)) + (1+\lambda_1)\min_{\theta\in\Theta}\frac{1}{T}\sum_{i=t}^{t+T} w_\theta^2(i)$$

for all $t \geqq d$ and T. This tells us why optimality can reasonably be expected to hold.

4. Stability in ideal, not necessarily positive real case. In this section we analyze the performance of the adaptive controller when it is applied to minimum phase ARMAX systems of *known* order. We do *not* make the usual positive-real assumption on the coefficients of the colored noise polynomial; in fact, we do not even assume any *stochastic* properties of the disturbance except for mean-square boundedness. Nevertheless we show that the adaptive controller proposed in the previous section mean-square *stabilizes* the system. (In the next section we show that stability result can be strengthened to one of *optimality* when a positive real condition is satisfied.)

We consider therefore the following ideal system:

$$A(q^{-1})y(t) = q^{-d}B(q^{-1})u(t) + C(q^{-1})w(t), \qquad t \geqq 1 \tag{4.1}$$

where

$$A(q^{-1}) = 1 + \sum_{i=1}^{n_A} a_i q^{-i}, \quad B(q^{-1}) = \sum_{i=0}^{n_B} b_i q^{-i}, \quad b_0 \neq 0, \quad C(q^{-1}) = 1 + \sum_{i=1}^{n_C} c_i q^{-i}.$$

Note that we assume the following:

(A4.i) Positive numbers λ_0, λ_1, delay d and reference output $y^m(t)$ are the same as that used in the adaptive controller (see § 2).

We only assume that the noise or disturbance $\{w(t)\}$ is mean-square bounded, i.e.,

(A4.ii) $$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} w^2(t) \leqq K < \infty \quad \text{a.s.}$$

Regarding the polynomials A, B, and C, we make the following assumptions.

(A4.iii) $B(z)$ has all zeros outside the closed unit disk.

Given $A(q^{-1})$, $B(q^{-1})$, and $C(q^{-1})$, there exist polynomials $S^*(q^{-1})$, $R^*(q^{-1})$, and $Q^*(q^{-1})$ so that

$$S^*(q^{-1})A(q^{-1}) + q^{-d}R^*(q^{-1})B(q^{-1}) = B(q^{-1}), \tag{4.2}$$

$$S^*(q^{-1}) = Q^*(q^{-1})B(q^{-1}). \tag{4.3}$$

We will assume that, with n_S and n_R corresponding to the choices in the adaptive controller, we have

(A4.iv) $S^*(q^{-1})$ is of degree n_S,

(A4.v) $R^*(q^{-1})$ is of degree n_R, i.e.,

$$S^*(q^{-1}) = \sum_{i=0}^{n_S} s_i^* q^{-i} \quad \text{and} \quad R^*(q^{-1}) = \sum_{i=0}^{n_R} r_i^* q^{-i}.$$

We define

$$\theta^* = (s_0^*, \cdots, s_{n_S}^*, r_0^*, \cdots, r_{n_R}^*, 0, \cdots, 0)^T, \tag{4.4}$$

and make the following assumptions:

(A4.vi) $(\theta^c-\theta^*)^T(\theta^c-\theta^*)\leqq K$;

(A4.vii) $s_0^*\geqq\sigma_0$, where s_0^* is the first component of θ^*.

With these assumptions in hand, we can proceed to our proof of stability. In what follows we denote the components of $\theta(t)$, the parameter estimate, by

$$\theta(t)=(s_0(t),\cdots,s_{n_S}(t),r_0(t),\cdots,r_{n_R}(t),-c_1(t),\cdots,-c_{n_C}(t))^T. \tag{4.5}$$

As we have observed at the end of § 3, we must first understand how the normalizing sequence $\rho(\cdot)$ is related to $e^2(t)$, the sum of the squares of $y(\cdot)$ and $u(\cdot)$.

LEMMA 4.1.

$$T\rho\leqq\rho(T)\leqq K_2\left(\sum_{t=1}^{T}y^{m^2}(t)+\sum_{t=1}^{T-1}(e^2(t)+w^2(t))\right)+T\rho \quad \textit{for some constant } K_2.$$

Proof. Let

$$\phi^r(t):=(u(t-1),\cdots,u(t-n_S),y(t),\cdots,y(t-n_R),y^m(t+d-1),\cdots,y^m\cdot(t+d-n_C))^T,$$

$$\theta^r(t):=(s_1(t),\cdots,s_{n_S}(t),r_0(t),\cdots,r_{n_R}(t),-c_1(t),\cdots,-c_{n_C}(t))^T.$$

Note that these "reduced" vectors are obtained by removing the first component from the vectors $\phi(t)$ and $\theta(t)$. From Lemma 3.3(iii) and assumption (A4.iii), we obtain that

$$\sum_{t=1}^{T}\|\phi^r(t-d)\|^2\leqq C_1\sum_{t=1}^{T-1}(y^{m^2}(t)+e^2(t)+w^2(t)), \tag{4.6}$$

for some constant C_1. From (2.1) we have

$$\rho(T)\leqq T\rho+\sum_{t=1}^{T}(u^2(t-d)+\|\phi^r(t-d)\|^2). \tag{4.7}$$

Using (2.11), (4.6), and Lemma 3.3(ii), we get

$$\begin{aligned}\sum_{t=1}^{T}u^2(t-d)&=\sum_{t=1}^{T-d}\left(\frac{y^m(t+d)-\theta^{rT}(t)\phi^r(t)}{s_0(t)}\right)^2\\ &\leqq\frac{2}{\sigma_0^2}\sum_{t=1}^{T}y^{m^2}(t)+\frac{2R^2}{\sigma_0^2}\sum_{t=d+1}^{T}\|\phi^r(t-d)\|^2 \qquad (4.8)\\ &\leqq C_2\sum_{t=1}^{T}(y^{m^2}(t)+e^2(t)+w^2(t)) \quad \text{for some constant } C_2.\end{aligned}$$

When we combine (4.6)-(4.8) the result follows. □

The following is a technical result that we use below.

LEMMA 4.2. *Let $v(t)\geqq 0$ be a sequence of positive real numbers for all $t\geqq 1$. If $1/T\sum_{t=1}^{T}v(t)\leqq V$, for all $T\geqq 1$, then*

(i) $$\sum_{t=q+1}^{q+k}\frac{v(t)}{t}\leqq V\left(1+\log\frac{q+k}{q}\right) \quad \textit{where } q\geqq 1;$$

(ii) $$\sum_{t=q+1}^{q+k}\frac{v(t)}{t^\alpha}\leqq\frac{V}{1-\alpha}(q+k)^{1-\alpha} \quad \textit{where } q\geqq 1,\quad 0\leqq\alpha<1.$$

Proof. Let $X(T)=(1/T)\sum_{t=1}^{T} v(t)$; then $v(t)=tX(t)-(t-1)X(t-1)$.

(i)
$$\sum_{t=q+1}^{q+k}\frac{v(t)}{t}=\sum_{t=q+1}^{q+k}X(t)-X(t-1)+\frac{1}{t}X(t-1)$$
$$=X(q+k)-X(q)+\sum_{t=q+1}^{q+k}\frac{X(t-1)}{t}$$
$$\leqq V\left(1+\sum_{t=q+1}^{q+k}\frac{1}{t}\right)\leqq V\left(1+\log\frac{q+k}{q}\right);$$

(ii)
$$\sum_{t=q+1}^{q+k}\frac{v(t)}{t^{\alpha}}=\sum_{t=q+1}^{q+k}\frac{t}{t^{\alpha}}X(t)-\frac{(t-1)}{t^{\alpha}}X(t-1)$$
$$\leqq (q+k)^{1-\alpha}X(q+k)+\sum_{t=q+1}^{q+k}(t-1)\left(\frac{1}{(t-1)^{\alpha}}-\frac{1}{t^{\alpha}}\right)X(t-1).$$

If $0\leqq\alpha<1$, then $t^{\alpha}-(t-1)^{\alpha}\leqq\alpha(t-1)^{\alpha-1}$. Therefore we have

$$\sum_{t=q+1}^{q+k}(t-1)\left(\frac{1}{(t-1)^{\alpha}}-\frac{1}{t^{\alpha}}\right)\leqq\alpha\sum_{t=q+1}^{q+k}t^{-\alpha}\leqq\frac{\alpha}{1-\alpha}(q+k)^{1-\alpha}.$$

Hence the result follows. □

LEMMA 4.3. *For any α, $0\leqq\alpha<1$, and with $V_\theta(\cdot)$ the sequence shown to be bounded in Lemma 3.3, there exists a constant C such that*

$$\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-d)-V_\theta(t))\leqq C(q+k)^{\alpha},\qquad q\geqq d\geqq 1.$$

Proof. The proof is by induction. Consider the case where $d=1$. Then,

$$\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-1)-V_\theta(t))\quad(\text{where } 0\leqq V_\theta(t)\leqq V_4 \text{ from Lemma 3.3})$$
$$\leqq q^{\alpha}V_4+\alpha V_4\sum_{t=q+1}^{q+k}(t-1)^{\alpha-1}\leqq 2V_4(q+k)^{\alpha}.$$

The induction is now on d. Suppose that for $i=1,\cdots,d-1$ there exist C_i such that

$$\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-i)-V_\theta(t))\leqq C_i(q+k)^{\alpha},\qquad q\geqq d.$$

Then, let us consider

$$\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-d)-V_\theta(t))$$
$$=\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-d)-V_\theta(t-d+1))+\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-d+1)-V_\theta(t))$$
$$\leqq\sum_{t=q+1}^{q+k}t^{\alpha}(V_\theta(t-d)-V_\theta(t-d+1))+C_{d-1}(q+k)^{\alpha}$$
$$\leqq 2(q+k)^{\alpha}(d-1)V_4+(2V_4+C_{d-1})(q+k)^{\alpha}=:C_d(q+k)^{\alpha}$$

and the induction is complete. □

We now reinterpret Lemma 3.3(v) to show that $\bar{e}(t)$ is small in the mean-square sense. This will then show that the operator $e(t)\to\sqrt{\rho(t)}$ can be considered small in the mean static gain operator, at least as far as norms are considered. Unfortunately, this property holds true only for time intervals where $\rho(t)$ is much larger than t.

LEMMA 4.4. *There exist almost surely finite random variables* $\tilde{L}$ *and* $\tilde{w}$ *such that for some given* $\varepsilon > 0$, $q+1 \leqq t \leqq q+k$, $q \geqq 1$, *if* $t/\rho(t) \leqq \varepsilon$ *then* $\sum_{t=q+1}^{q+k} \bar{e}^2(t) \leqq \tilde{L} + \varepsilon\tilde{w}(1+\lambda_1)\log(q+k)/q$.

Proof. From Lemma 3.3(v) and Lemma 3.2, we have

$$\sum_{t=q+1}^{q+k} \bar{e}^2(t) \leqq C_0 + (1+\lambda_1) \min_{\theta\in\Theta} \left(\sum_{t=q+1}^{q+k} \bar{w}_\theta^2(t) \right) \tag{4.9}$$

for some constant C_0. By the definition of θ^*, we have

$$w_{\theta^*}(t) = Q^*(q^{-1})C(q^{-1})w(t). \tag{4.10}$$

Because $w(t)$ is almost surely mean-square bounded, i.e., $\limsup_T (1/T)\sum_{t=1}^{T} w^2(t) < \infty$ a.s., $Q^*(q^{-1})$ and $C(q^{-1})$ are polynomials, from (4.10) we see that there exists an almost surely finite random variable $\tilde{w}$ such that

$$\sup \frac{1}{T} \sum_{t=1}^{T} w_{\theta^*}^2(t) \leqq \tilde{w} \quad \text{a.s.} \tag{4.11}$$

If $t/\rho(t) \leqq \varepsilon$, then $\bar{w}_{\theta^*}^2(t) \leqq \varepsilon(w_{\theta^*}^2(t)/t)$ for $t \in [q+1, q+k]$, $q \geqq 1$. Combining this inequality, Lemma 4.2(i), and (4.11), we have

$$\sum_{t=q+1}^{q+k} \bar{e}^2(t) \leqq C_0 + (1+\lambda_1)\varepsilon\tilde{w}\left(1+\log\frac{q+k}{q}\right) = \tilde{L} + \varepsilon\tilde{w}(1+\lambda_1)\log\frac{q+k}{q}. \qquad \square$$

With Lemmas 4.1 and 4.4 now established, we are in a position to "close the loop." To do so we need an appropriate version of the small gain theorem given in the next result.

LEMMA 4.5 (Bellman-Gronwall Lemma). *If* $\rho(T) \leqq \tilde{M}_4 T + M_2\rho(T_0) + \gamma\sum_{t=T_0+1}^{T-1} \bar{e}^2(t)\rho(t)$, *then*

$$\rho(T) \leqq \tilde{M}_4 T + M_2\rho(T_0) \prod_{t=T_0+1}^{T-1} (1+\gamma\bar{e}^2(t)) + \gamma\tilde{M}_4 \sum_{t=T_0+1}^{T-1} t\bar{e}^2(t) \prod_{i=t+1}^{T-1} (1+\gamma\bar{e}^2(i))$$

for some positive constant M_2 *and some positive random variable* $\tilde{M}_4$.

Proof. The proof uses mathematical induction and we provide a sketch. For $T = T_0+1$ statement is obviously true. Suppose that the statement is true for $T_0+1 \leqq T \leqq T_1$, then $\rho(T_1+1) \leqq \tilde{M}_4(T_1+1) + M_2\rho(T_0)X_1 + \gamma\tilde{M}_4 X_2$, where

$$X_1 := 1 + \gamma \sum_{t=T_0+1}^{T_1} \bar{e}^2(t) \prod_{j=T_0+1}^{t-1} (1+\gamma\bar{e}^2(j)) = \prod_{j=T_0+1}^{T_1} (1+\gamma\bar{e}^2(j))$$

and

$$\begin{aligned} X_2 &:= \sum_{t=T_0+1}^{T_1} t\bar{e}^2(t) + \gamma \sum_{t=T_0+1}^{T_1} \bar{e}^2(t) \sum_{j=T_0+1}^{t-1} j\bar{e}^2(j) \prod_{i=j+1}^{t-1} (1+\gamma\bar{e}^2(i)) \\ &= \sum_{t=T_0+1}^{T_1} t\bar{e}^2(t) \prod_{i=t+1}^{T_1} (1+\gamma\bar{e}^2(i)). \end{aligned}$$

$\square$

We now show that the adaptive controller mean square stabilizes the system under the assumptions stated at the beginning of this section. Note that we are *not* assuming a positive real condition on the noise.

THEOREM 4.6. *For system* (4.1), *subject to the assumptions* (A4.i)-(A4.vii), *our algorithm ensures that*:

(i) $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} y^2(t) < \infty$ *a.s.*,

(ii) $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u^2(t) < \infty$ *a.s.*

Proof. For $\tilde{w}$ (given by (4.11)), K_2 (given by Lemma 4.1), and λ_1, there exist random variables $\tilde{\varepsilon}$ and $\tilde{\alpha}$ such that

$$\tilde{\alpha} = 1 - K_2\tilde{\varepsilon}\tilde{w}(1+\lambda_1) \quad \text{and} \quad 0 < \tilde{\alpha} < 1 \quad \text{a.s.} \tag{4.12}$$

Suppose now that there exists a time interval $(T_0, T_1]$ such that

$$\frac{T}{\rho(T)} \leqq \tilde{\varepsilon} \quad \text{for all } T \in (T_0, T_1] \quad \text{and} \quad T_0/\rho(T_0) > \tilde{\varepsilon}$$

where T_1 may be infinite. (Note that if such an interval does not exist, then we are done.)

Because $y^m(t)$ is uniformly bounded and $w(t)$ is almost surely mean-square bounded, by means of Lemma 4.1, there exists $\tilde{M}_1$ such that $\rho(T) \leqq \tilde{M}_1 T + K_2 \sum_{t=1}^{T-1} e^2(t)$ almost surely.

Since $e(t) = (\theta^* - \theta(t-d))^T \phi(t-d) + w_{\theta^*}(t)$, from Lemma 3.1 and (4.11) there exist M_2 and $\tilde{M}_3$ such that $\sum_{t=1}^{T_0} e^2(t) \leqq M_2\rho(T_0) + \tilde{M}_3 T_0$.

Using this inequality and the Bellman-Gronwall Lemma, we have

$$\rho(T) \leqq \tilde{M}_4 T + M_5\rho(T_0) \prod_{t=T_0+1}^{T-1} (1 + K_2\bar{e}^2(t)) + K_2\tilde{M}_4 \sum_{t=T_0+1}^{T-1} t\bar{e}^2(t) \prod_{i=t+1}^{T-1} (1 + K_2\bar{e}^2(i)).$$

From Lemma 4.4 and (4.12), we have $\prod_{t=q+1}^{q+k} (1 + K_2\bar{e}^2(t)) \leqq e^{K_2\tilde{L}((q+k)/q)^{1-\tilde{\alpha}}}$. Therefore there exists an almost surely finite random variable $\tilde{M}_6$ such that

$$\rho(T) \leqq \tilde{M}_6\left[T + \rho(T_0)\left(\frac{T-1}{T_0}\right)^{1-\tilde{\alpha}} + (T-1)^{1-\tilde{\alpha}} \sum_{t=T_0+1}^{T-1} t^{\tilde{\alpha}}\bar{e}^2(t) \right]. \tag{4.13}$$

Choosing $\theta = \theta^*$ in Lemma 3.2, we have $(1/(1+\lambda_1))\bar{e}^2(t) \leqq (V_{\theta^*}(t-d) - V_{\theta^*}(t)) + \bar{w}^2_{\theta^*}(t)$. Hence we get

$$\sum_{t=T_0+1}^{T-1} t^{\tilde{\alpha}}\bar{e}^2(t) \leqq (1+\lambda_1) \sum_{t=T_0+1}^{T-1} t^{\tilde{\alpha}}(V_{\theta^*}(t-d) - V_{\theta^*}(t) + \bar{w}^2_{\theta^*}(t)).$$

From Lemmas 4.3 and 4.2(ii), we have

$$\sum_{t=T_0+1}^{T-1} t^{\tilde{\alpha}}\bar{e}^2(t) \leqq C_1(T-1)^{\tilde{\alpha}} + (1+\lambda_1)\tilde{\varepsilon} \sum_{t=T_0+1}^{T-1} \frac{w^2_{\theta^*}(t)}{t^{1-\tilde{\alpha}}} \leqq \tilde{M}_7(T-1)^{\tilde{\alpha}},$$

for some $\tilde{M}_7$. We can rewrite (4.13) as $\rho(T)/T \leqq \tilde{M}_6(1 + 1/\tilde{\varepsilon} + \tilde{M}_7)$. Hence there exists a random variable $\tilde{\varepsilon}_1$ such that

$$\frac{T}{\rho(T)} \geqq \tilde{\varepsilon}_1 > 0. \tag{4.14}$$

From Lemma 3.1(iii), we know that $1/\tilde{\varepsilon} > (1/T)\sum_{t=1}^{T-d} y^2(t)$. This implies that $\limsup_{T\to\infty} 1/T\sum_{t=1}^{T} y^2(t) < \infty$ a.s. Similarly, $\limsup_{T\to\infty} (1/T)\sum_{t=1}^{T} u^2(t) < \infty$ a.s. □

5. Optimality in the ideal, positive real case. Now we turn attention to the so-called *ideal case*, where the noise *satisfies a positive real condition*, and show that the preceding

stability results can be improved to prove that the sample mean-square variance of the output error is actually *optimal.* Also, in the case of *regulation,* we prove that the adaptive controller *self-tunes in a Cesaro sense* to a minimum variance regulator.

Given the system in the previous section, let us suppose that the polynomials

$$S(q^{-1}) := \sum_{i=0}^{n_S} s_i q^{-i} \quad \text{and} \quad R(q^{-1}) := \sum_{i=0}^{n_R} r_i q^{-i}$$

satisfy the equations

$$S(q^{-1})A(q^{-1}) + q^{-d}R(q^{-1})B(q^{-1}) = C(q^{-1})B(q^{-1}), \tag{5.1}$$

$$S(q^{-1}) = Q(q^{-1})B(q^{-1}). \tag{5.2}$$

Then we can define $\theta^0(t) := (s_0, \cdots, s_{n_S}, r_0, \cdots, r_{n_R}, -c_1, \cdots, -c_{n_C})^T$.

Regarding the noise $\{w(t)\}$ we assume the following:

(A5.i) It is a martingale difference sequence on a probability space (Ω, F, P).

Specifically, denoting by F_t the sub-σ-algebra generated by the observation up to and including time t. We assume that:

(A5.ii) $E\{w(t)|F_{t-1}\} = 0$ a.s.;

(A5.iii) $E\{w^2(t)|F_{t-1}\} = \sigma^2$ a.s.;

(A5.iv) $\sup_t E\{|w(t)|^{2+\delta}|F_{t-1}\} < \infty$ a.s. for some $\delta > 0$.

Next, let $v(t) := Q(q^{-1})w(t)$; then

$$E\{v^2(t+d)|F_t\} = \sigma^2 \sum_{i=0}^{d-1} q_i^2 =: v^2 \quad \text{a.s.} \quad \text{where } Q(q^{-1}) := \sum_{i=0}^{d-1} q_i q^{-i}. \tag{5.3}$$

Clearly the minimum tracking variance is v^2 (see Kumar and Varaiya [12]). We now show that our adaptive controller achieves this optimal tracking performance.

THEOREM 5.1. *Suppose that the system* (4.1) *satisfies assumptions* (A4.i), (A4.iii)-(A4.vii), *and* (A5.i)-(A5.iv). *Furthermore, assume the positive realness condition* $\sup_\omega |C(e^{i\omega}) - 1| < 1/\sqrt{1+\lambda_1}$ *and also that* $\theta^0 \in \Theta$. *Then* $\lim_{T\to\infty} 1/T \sum_{t=d}^{T+d} (y(t) - y^m(t))^2 = v^2$ *almost surely.*

Proof. From (5.1) and (4.1), it is easy to see that $(e(t) - v(t))$ is F_{t-d}-measurable. Now let

$$z(t-d) := e(t) - v(t), \tag{5.4}$$

$$b(t) := (\theta^0 - \theta(t))^T \phi(t), \tag{5.5}$$

$$h(t) := b(t) - z(t); \tag{5.6}$$

then it is easy to see that $C(q^{-1})z(t) = b(t)$. Hence $h(t) = (C(q^{-1}) - 1)z(t)$.

Because $C(e^{i\omega})$ is strictly inside the circle with center 1, and radius $1/\sqrt{1+\lambda_1}$, there exists a positive ε such that $\sum_{j=1}^{t} (z^2(j)/(1+\lambda_1) - h^2(j)) \geqq \varepsilon \sum_{j=1}^{t} z^2(j)$ for all t.

Let us define a function

$$S(t) := \sum_{j=d+1}^{t} \left(\frac{z^2(j-d)}{1+\lambda_1} - h^2(j-d) - \varepsilon z^2(j-d) \right), \qquad t \geqq d+1 \tag{5.7}$$

with $S(d) := 0$. Obviously, $S(t) \geqq 0$ for $t \geqq d$ and

$$S(t) - S(t-1) = \frac{z^2(t-d)}{1+\lambda_1} - h^2(t-d) - \varepsilon z^2(t-d) \quad \text{for } t \geqq d+1.$$

Since $w_{\theta^0}(t) = e(t) - b(t-d)$, from Lemma 3.2 and (5.4) we get

$$\begin{aligned} V_{\theta^0}(t) &\leqq V_{\theta^0}(t-d) + (\bar{e}(t) - \bar{b}(t-d))^2 - g(t)\bar{e}^2(t) \\ &= V_{\theta^0}(t-d) + (\bar{z}(t-d) + \bar{v}(t))^2 - 2(\bar{z}(t-d) + \bar{v}(t))\bar{b}(t-d) + \bar{b}^2(t-d) \\ &\quad - g(t)(\bar{z}(t-d) + \bar{v}(t))^2 \end{aligned}$$

where $\bar{b}(t-d) := b(t-d)/\rho^{1/2}(t)$, $\bar{v}(t) := v(t)/\rho^{1/2}(t)$, and $\bar{z}(t-d) := z(t-d)/\rho^{1/2}(t)$.

Taking the conditional expectation and using Lemma 3.1(i),

$$\begin{aligned} E\{V_{\theta^0}(t) \mid F_{t-d}\} &\leqq V_{\theta^0}(t-d) + (1-g(t))\frac{v^2}{\rho(t)} + (\bar{z}(t-d) - \bar{b}(t-d))^2 - \frac{1}{1+\lambda_1}\bar{z}^2(t-d) \\ &\leqq V_{\theta^0}(t-d) + (1-g(t))\frac{v^2}{\rho(t)} + \frac{S(t-1) - S(t)}{\rho(t)} - \varepsilon\bar{z}^2(t-d). \end{aligned}$$

However,

$$\frac{S(t-1)}{\rho(t)} \leqq \frac{S(t-1)}{\rho(t-1)}, \qquad \frac{S(d)}{\rho(d)} \leqq \frac{M_1}{d\rho},$$

and so,

$$\sum_{t=d+1}^{T} E\{V_{\theta^0}(t) \mid F_{t-d}\} \leqq \sum_{t=d+1}^{T} V_{\theta^0}(t-d) + v^2 \sum_{t=d+1}^{T} \frac{1-g(t)}{\rho(t)} + \frac{M_1}{d\rho} - \varepsilon \sum_{t=d+1}^{T} \bar{z}^2(t-d).$$

Because

$$\begin{aligned} \frac{1-g(t)}{\rho(t)} &= \frac{\bar{\phi}^T(t-d)F(t-d)\bar{\phi}(t-d)}{\rho(t)(1+\bar{\phi}^T(t-d)F(t-d)\bar{\phi}(t-d))} \\ &\leqq \frac{\lambda_1\phi^T(t-d)\phi(t-d)}{\rho(t)(\rho(t)+\lambda_1\phi^T(t-d)\phi(t-d))} \\ &\leqq \lambda_1\frac{\rho(t)-\rho(t-1)}{\rho^2(t)} \leqq \lambda_1\left(\frac{1}{\rho(t-1)} - \frac{1}{\rho(t)}\right). \end{aligned}$$

This implies that $\sum_{t=d+1}^{T}(1-g(t))/\rho(t) \leqq \lambda_1/d\rho$.

Taking unconditional expectation, and noting that $V_{\theta^0}(t)$ is bounded (surely), there exists M_2 such that $\varepsilon E\{\sum_{t=d+1}^{T}\bar{z}^2(t-d)\} \leqq M_2$. Hence

$$\sum_{t=d+1}^{\infty} \frac{(e(t)-v(t))^2}{\rho(t)} < \infty \quad \text{a.s.} \tag{5.8}$$

From Kronecker's lemma and (4.14), we get

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(e(t)-v(t))^2 = 0 \quad \text{a.s.} \tag{5.9}$$

Hence $E\{(y(t)-y^m(t))^2 \mid F_{t-d}\} = E\{(e(t)-v(t)+v(t))^2 \mid F_{t-d}\} = (e(t)-v(t))^2 + v^2$.

Now, continuing as in Lemma 7 of Becker, Kumar, and Wei [3], we get the desired result. □

It should be noted that as $(1+\lambda_1)$ increases, the *speed* of adaptation is increased. However, the condition $\sup_\omega |C(e^{i\omega}) - 1| < 1/\sqrt{1+\lambda_1}$ then becomes more stringent, requiring that the noise be even closer to pure white noise. Hence we see that λ_1 allows a tradeoff between the rate of parameter convergence and the tolerance of the algorithm to colored noise.

In fact, we can even prove that the adaptive regulator self-tunes in a Cesaro sense to the set of optimum minimum variance regulators. To exhibit this result, we concentrate temporarily on the regulation problem. In this case,

$$y^m(t)=0 \quad \text{for every } t,$$

$$\theta^T(t)=(\theta_1(t),\cdots,\theta_{n_S+1}(t),\theta_{n_S+2}(t),\cdots,\theta_{n_S+n_R+2}(t)),$$

$$\phi^T(t)=(u(t),\cdots,u(t-n_S),y(t),\cdots,y(t-n_R))$$

and (2.11) can be rewritten as

$$\theta^T(t)\phi(t)=0. \tag{5.10}$$

Let us define $R'(q^{-1},\theta(t)):=\sum_{i=0}^{n_R}\theta_{n_S+2+i}(t)q^{-i}$ and $S'(q^{-1},\theta(t)):=\sum_{i=0}^{n_S}\theta_{i+1}(t)q^{-i}$. Then from (5.10), we have $u(t)=-(R'(q^{-1},\theta(t))/S'(q^{-1},\theta(t)))y(t)$.

Note that $D:=\{\theta\,|\,R'(q^{-1},\theta)S(q^{-1})=S'(q^{-1},\theta)R(q^{-1})\}$ is the set of parameters that yield a minimum variance regulator. We now have the following result on self-tuning in a Cesaro sense.

THEOREM 5.2. *For every open set* $O\supset D$, $\lim_{T\to\infty}1/T\sum_{t=1}^{T}1(\theta(t)\in O)=1$ *almost surely, where* $1(\cdot)$ *is the indicator function.*

Proof. Because $z(t)=y(t+d)-Q(q^{-1})w(t+d)=E\{y(t+d)\,|\,F_t\}$, from (5.9) we know that (14.i) in Becker, Kumar, and Wei [3] is true. From Lemma 3.3(iv), we have

$$\|\theta(t)-\theta(t-d)\|^2\leqq 2\lambda_i\left(1+\frac{\sqrt{\lambda_1}}{\sqrt{\lambda_0}}\right)^2\left[\frac{(e(t)-v(t))^2}{\rho(t)}+\frac{v^2(t)}{\rho(t)}\right].$$

It is easy to see that $\{X(t):=\sum_{i=1}^{t}(v^2(i)-v)/i;\,F_t\}$ is a martingale. Due to (A5.iv) $X(t)$ converges and so on $(v^2(t)-v)/t\to 0$ almost surely and this implies $v^2(t)/t\to 0$ almost surely. Because $1/\rho$ is the upper bound of $t/\rho(t)$ for every $t>0$, so $v^2(t)/\rho(t)=(v^2(t)/t)t/\rho(t)\to 0$ almost surely. Combining with (5.8), we get $\|\theta(t)-\theta(t-d)\|^2\to 0$ almost surely. Therefore (14.ii) in [3] is true.

From Lemma 3.3(ii), (5.10), Theorem 4.6(ii), and Lemma 3.3(i), we can see that (14.iii)-(14.vi) in [3] are true. Hence our result follows from Theorem 19(ii) in [3]. □

6. Robustness of optimal adaptive controller. Having proved in the previous section that our adaptive controller yields *optimal performance* for ideal systems, we now turn in this section to proving that the preceding adaptive controller is *robust.* This means that if mean-square stability holds for an ideal system Π_0 (and it does, as we have shown), it will continue to hold for all systems in an *open neighborhood* of Π_0. For this to make sense, we need to define a topology on the set of linear systems. We will consider the *graph topology* (see Vidyasagar [26]) and show that the adaptive controller applied to systems in a graph topological neighborhood retains mean-square stability. Furthermore, we will also give lower bounds on the size of these graph topological neighborhoods.

Let Γ be the set of proper rational functions $F(q)$ whose poles are all in the open unit disk. Γ is equipped with the norm $\gamma(\cdot)$, defined as $\gamma(F):=\sup_{|q|=1}|F(q)|$, for all $F\in\Gamma$. For a sequence $x(t)$, we define its l_2-norm as $\|x\|_T^2:=\sum_{t=1}^{T}x^2(t)$.

We will prove that our adaptive controller stabilizes nonideal systems Π if they satisfy the following assumptions:

(A6.i) Let the system Π be described by the equation

$$A(q)y(t)=B(q)u(t-1)+C(q)w(t), \qquad t\geqq 1 \tag{6.1}$$

where A, B, $C \in \Gamma$, A, B are coprime, B/A is a proper rational function, $A(\infty) = 1 = C(\infty)$, and the noise is a stochastic process that satisfies merely $\sup_T (1/T \sum_{t=1}^{T} w^2(t))^{1/2} \leqq V$, where V is a *deterministic* finite number. (This clearly holds if, for example, the noise is bounded.) We also assume that $|y^m(t)| \leqq M$ for all $t > 0$ and $y^m(t) = u(t) = y(t) = w(t) = 0$ for all $t \leqq 0$.

Because $B(q)$ is an analytic function outside the unit disk, we can write a Laurent series $B(q) = \sum_{i=0}^{\infty} h_i q^{-i}$ and, for $d \geqq 2$, set $P(q^{-1})$ equal to $\sum_{i=0}^{d-2} h_i q^{-i}$; otherwise it equals zero, and $D(q) := \sum_{i=0}^{\infty} h_{i+d-1} q^{-i}$.

It is easy to see that

$$B(q) = P(q^{-1}) + q^{1-d} D(q). \tag{6.2}$$

Note that for the ideal system Π_0 (as in § 5), $P_0(q^{-1}) = 0$ and $D_0(q) = B_0(q^{-1}) \in \Gamma$. Because Π_0 is minimum phase, $D_0(q)$ is strictly stably invertible. Motivated by this, we assume the following:

(A6.ii) $D(q)$ is an invertible element of Γ (i.e., $D(q)$ and $D^{-1}(q)$ belong to Γ, or we can say $D(q)$ is a *unit* of Γ).

For a system Π, we define

$$T(q) = (A(q), \cdots, q^{-n_S} A(q), q^{-1} B(q), \cdots, q^{-(n_R+1)} B(q), 0, \cdots, 0)^T.$$

With $D(q)$ a unit of Γ, for every θ we can define a new element of Γ by

$$H_\theta(q) := 1 - D^{-1}(q) T^T(q) \theta. \tag{6.3}$$

Clearly $\gamma(H_\theta)$ is a continuous function of θ. Hence we can choose $\tilde{\theta} \in \Theta$ so that $\gamma(H_{\tilde{\theta}}) \leqq \gamma(H_\theta)$ for all $\theta \in \Theta$. Next we assume the following:

(A6.iii) $\gamma(H_{\tilde{\theta}}) < \gamma_h$, where $\gamma_h = 1/\gamma_4 = 1/\sqrt{1+\lambda_1}$;

(A6.iv) $\gamma(P) < (\gamma_h - \gamma(H_{\tilde{\theta}})/(\gamma(D^{-1})(k_1 \gamma(D^{-1}) \gamma(A) + k_2 + k_3(\gamma_h - \gamma(H_{\tilde{\theta}}))) \gamma_3^{d-1}))$, where k_1, k_2, and k_3 are strictly positive constants given in the Table 1 in the Appendix (as is γ_3 also).

We illustrate these assumptions by the following two examples.

Example 1. Consider the adaptive controller with $n_S = 0$, $n_R = 0$, $n_C = 0$, i.e., $\phi(t) = (u(t), y(t))$. We now examine the above assumptions by allowing only one parameter to vary. Consider $\theta = (1, r)$, which denotes that the adaptive controller is associated with the idealized plant

$$(1 + rq^{-1}) y(t) = u(t-1) + w(t).$$

However, suppose that the true plant is given by

$$A(q) = 1 + a_1 q^{-1} + a_2 q^{-2}, \quad B(q) = 1, \quad C(q) = 1.$$

Then straightforward computations give

$$T(q) = (1 + a_1 q^{-1} + a_2 q^{-2}, q^{-1})^T, \qquad D(q) = 1,$$
$$H_\theta(q) = -q^{-1}(a_1 + r + a_2 q^{-1})$$

and it follows that

$$\gamma(H_\theta) = \sup_{0 \leqq \alpha \leqq 2\pi} \sqrt{(a_1+r)^2 + 2a_2(a_1+r)\cos\alpha + a_2^2} = |a_1 + r| + |a_2|.$$

Therefore we get $\tilde{\theta} = (1, -a_1)^T$, $H_{\tilde{\theta}} = -a_2 q^{-2}$, and $\gamma(H_{\tilde{\theta}}) = |a_2|$.

Hence, for this problem, with the expression of γ_h, assumption (A6.iii) is just equivalent to $|a_2| < 1/\sqrt{1+\lambda_1}$.

Note that all the other assumptions are satisfied. Thus by this inequality we see that λ_1 also allows a tradeoff between the rate of parameter convergence and the size of the allowed value of a_2 (see also the last comment of § 5).

Example 2. To illustrate that our assumptions do not require the system to be of minimum phase, we now consider $n_S = 1$, $n_R = 0$, $n_C = 0$, $d = 2$, i.e., $\phi(t) = (u(t), u(t-1), y(t))$. We now study the assumptions by the variation of two parameters, so suppose that $\theta = (1, s, r)^T$; this clearly corresponds to an adaptive controller for an idealized plant:

$$(1+aq^{-1})y(t) = q^{-1}u(t-1) + w(t).$$

However, suppose that the true plant is $(1+aq^{-1})y(t) = (b+q^{-1})u(t-1) + w(t)$. Then we have

$$T(q) = (1+aq^{-1}, q^{-1}(1+aq^{-1}), (b+q^{-1})q^{-1}), \qquad D(q) = 1,$$

$$P(q^{-1}) = b, \qquad H_\theta(q) = -q^{-1}(a+s+rb+(as+r)q^{-1}).$$

Therefore we get

$$\gamma(H_\theta) = |a+s+rb| + |as+r|, \quad \gamma(H_{\tilde\theta}) = 0, \quad \tilde\theta = \left(1, -\frac{a}{1-ab}, \frac{a^2}{1-ab}\right).$$

Hence all the assumptions are satisfied if (A6.iv) holds, i.e.,

$$|b| < \frac{\gamma_h}{k_1(1+|a|)+k_2+k_3\gamma_h\gamma_3},$$

which reduces to

$$|b|\left\{\sqrt{2(1+\lambda_1)}R(1+|a|)+R\sqrt{1+\lambda_1}+2\sqrt{2}\left\{\left(1+\frac{2R^2}{\sigma_0^2}\right)\left[\frac{\sup|w(t)|}{\sqrt{\rho}}+2(|a|+|b|)+2\right]^2 + 3+\frac{\sup y^{m^2}}{\rho}\left[1+\frac{2}{\sigma_0^2}(1+R^2)\right]+\frac{2R^2}{\sigma_0^2}\right\}^{1/2}\right\} < 1.$$

Note that the actual plant has a zero at $-1/b$. Thus we see that if the plant is nonminimum phase, then we can model the unstable zeros by delays, provided these zeros are *large* enough. We notice that by reducing the size of the parameter domain (i.e., by decreasing R and increasing σ_0), we allow smaller unstable zeros. This is a manifestation of the well-known fact that high gains may cause problems in the presence of unmodeled dynamics.

We see also that the threshold for the unmodeled unstable zero depends on the l_∞-norm of the forcing signals w and y^m of the closed-loop system. This is a manifestation of its nonlinear nature. However, since these norms are divided by $\sqrt{\rho}$, we can overcome this difficulty by choosing the threshold ρ in (2.1) proportional to the square of these norms.

We consider a graph topology constructed from the set Γ. All the properties of [26] can be rederived here. Specifically, this topology is the weakest one such that feedback stability is robust and closed-loop transfer functions are continuous (with respect to the "sup" norm). Since this topology on the collection of systems Π follows from the topology on Γ^3, our robustness result follows from the following theorem.

THEOREM 6.1. *The set of* (A, B, C) *satisfying assuptions* (A6.ii)-(A6.iv) *is open.*

Proof. The set

$$\left\{(\bar{A}, \bar{B}, \bar{C}, F, H, G) \,|\, \bar{A}, \bar{B}, \bar{C}, F, H, G \in \Gamma,\ F \in U,\ \gamma(H) < \gamma_h,\right.$$

$$\left.\gamma(G) < \frac{\gamma_h - \gamma(H)}{\gamma(F)(k_1\gamma(F)\gamma(\bar{A}) + k_2 + k_3(\gamma_h - \gamma(H)))\gamma_3^{d-1}(\bar{A}, \bar{B}, \bar{C})}\right\}$$

is an open set of Γ^6 where U denotes the set of units of Γ. This holds since U is an open subset of Γ, and the mapping $F^{-1} \to F$ is continuous on U (see [26]).

Let us prove that the mapping $(A, B, C) \to P$ is continuous. Using the Cauchy-Schwarz inequality, and Parseval's theorem (see [4]) we have $\gamma(P) \leqq \sum_{i=0}^{d-2} |h_i| \leqq \sqrt{d-1}\gamma(B)$. Since the mapping $(A, B, C) \to P$ is linear, and as we have just shown, also bounded, this proves that it is continuous. This implies that the mapping $(A, B, C) \to D$ is continuous, and $(A, B, C) \to H_\theta$ is continuous, for any fixed θ. Hence, for any fixed θ, the mapping $(A, B, C) \to (D, H_\theta, P)$ is continuous.

Therefore the set

$$\psi_\theta := \left\{(A, B, C) \,|\, D \in U,\ \gamma(H_\theta) < \gamma_h,\right.$$

$$\left.\gamma(P) < \frac{\gamma_h - \gamma(H_\theta)}{\gamma(D^{-1})(k_1\gamma(D^{-1})\gamma(A) + k_2 + k_3(\gamma_h - \gamma(H_\theta)))\gamma_3^{d-1}(A, B, C)}\right\}$$

is open, and therefore $\bigcup_{\theta \in \Theta} \psi_\theta$ is also open. The result follows. □

Before showing the proof of Theorem 6.8, the main robustness theorem, we need some results. As Lemma 3.3 shows, it is sufficient to prove that the operator $e(t) \to \sqrt{\rho(t)}$ has a finite gain and the operator $e(t) \to w_{\tilde{\theta}}(t)$ has a gain bounded by $1/\sqrt{1+\lambda_1}$. In what follows, we use a number of positive constants α_i, β_i, γ, δ_i, V_i, and k_i, given in Table 1 in the Appendix, that depend on $\gamma(A)$, $\gamma(B)$, $\gamma(C)$, M, V, K, R, n_C, n_S, n_R, μ, d, ρ, λ_0, λ_1, and σ_0.

LEMMA 6.2.

(i) $\rho(t) \geqq \rho(t-1)$;

(ii) $\|\phi\|_T^2 \leqq \rho(T+d) \leqq \|\phi\|_T^2 + T\rho + V_1$;

(iii) $\frac{1}{2}(\|u\|_t + \|y\|_t) \leqq \|\phi\|_t \leqq \gamma_1\|u\|_t + \gamma_2\|y\|_t + \sqrt{t}\alpha_1$;

(iv) $\|w\|_T \leqq \sqrt{T}V \leqq (V/\sqrt{\rho})\rho^{1/2}(T)$.

Proof. Formulae (i) and (iv) are immediate.

(ii) Since $\rho(t+d) - \rho(t+d-1) \leqq \rho + \|\phi(t)\|^2$, we have $\sum_{t=1}^{T}(\rho(t+d) - \rho(t+d-1)) \leqq T\rho + \|\phi\|_T^2$. Choosing $V_1 := \rho(d)$, we find that $\rho(T+d) \leqq T\rho + \|\phi\|_T^2 + V_1$.

(iii) The left-hand inequality is obvious; for the right-hand side,

$$\|\phi\|_t^2 \leqq (1+n_S)\|u\|_t^2 + (1+n_R)\|y\|_t^2 + n_C t M^2.$$

Now choosing $\gamma_1 := \sqrt{1+n_S}$, $\gamma_2 := \sqrt{1+n_R}$, and $\alpha_1 := M\sqrt{n_C}$, we get the result. □

LEMMA 6.3. $\rho(t+1) \leqq \gamma_3^2\rho(t)$.

Proof.

$$|y(t+1)| \leqq w(t+1)| + \gamma(B)\|u\|_t + \gamma(A-1)\|y\|_t + \gamma(C-1)\|w\|_t. \tag{6.4}$$

Define $\phi'(t) := (u(t-1), \cdots, u(t-n_S), y(t), \cdots, y(t-n_R), y^m(t+d-1), \cdots, y^m(t+d-n_C))^T$ and $\theta'(t) := (s_1(t), \cdots, s_{n_S}(t), r_0(t), \cdots, r_{n_R}(t), -c_1(t), \cdots, -c_{n_C}(t))^T$. Then the control law becomes

$$u(t) = \frac{y^m(t+d) - \theta^{rT}(t)\phi^r(t)}{s_0(t)}. \tag{6.5}$$

By Lemma 6.2(ii) we have

$$\|\phi^r(t)\|^2 \leqq \rho(t+d-1)+y^2(t)+M^2. \tag{6.6}$$

Putting (6.5) and (6.6) together, we have

$$u^2(t) \leqq \frac{2}{\sigma_0^2}(M^2+R^2(\rho(t+d-1)+y^2(t)+M^2)). \tag{6.7}$$

Next, with Lemma 6.2(ii), we have

$$\|\phi(t-d+1)\|^2 \leqq \rho(t)+u^2(t-d+1)+y^2(t-d+1)+M^2. \tag{6.8}$$

From (6.4) and Lemma 6.2, we have

$$\begin{aligned} y^2(t-d+1) &\leqq [|w(t-d+1)|+(\gamma(B)+\gamma(A-1))(\|u\|_{t-d}+\|y\|_{t-d}) \\ &\qquad +\gamma(C-1)\|w\|_{t-d}]^2 \qquad (6.9) \\ &\leqq \left(\frac{V}{\sqrt{\rho}}+2(\gamma(B)+\gamma(A-1))+\frac{\gamma(C-1)}{\sqrt{\rho}}V\right)^2\rho(t). \end{aligned}$$

Combining (6.7)-(6.9), we have

$$\begin{aligned} \rho(t+1) &\leqq \rho(t)+\rho+\|\phi(t-d+1)\|^2 \\ &\leqq \left(2+\frac{2}{\sigma_0^2}R^2\right)\rho(t)+\left(1+\frac{2}{\sigma_0^2}R^2\right)y^2(t-d+1)+M^2+\rho+\frac{2}{\sigma_0^2}M^2(1+R^2) \\ &\leqq \left(1+\frac{2}{\sigma_0^2}R^2\right)\left[\frac{V}{\sqrt{\rho}}+2(\gamma(B)+\gamma(A-1))+\gamma\frac{(C-1)}{\sqrt{\rho}}V\right]^2\rho(t) \\ &\quad +\left[2+\frac{2}{\sigma_0^2}R^2+\frac{M^2}{\rho}+1+\frac{2}{\rho\sigma_0^2}M^2(1+R^2)\right]\rho(t) \\ &=: \gamma_3^2\rho(t). \end{aligned}$$

□

LEMMA 6.4.

$$\|u\|_{t-d} \leqq \gamma(D^{-1})[\gamma(A)\|y\|_t+\gamma(C)\|w\|_t+\gamma(P)\|u\|_{t-1}].$$

Proof. Using (6.2), we have $D(q)u(t-d)=A(q)y(t)-C(q)w(t)-P(q^{-1})u(t-1)$, that is, $u(t-d)=D^{-1}(q)\{A(q)y(t)-C(q)w(t)-P(q^{-1})u(t-1)\}$. The result follows. □

The next lemma is immediate.

LEMMA 6.5.

(i) $\|y\|_t \leqq \|y^m\|_t+\|e\|_t$;

(ii) $\|\phi(t)\| \leqq \alpha_1$ *for* $t \leqq 0$, *where* $\alpha_1 = M\sqrt{n_C}$.

We therefore see from Lemmas 6.2, 6.4, and 6.5 that the operator $e(t) \to \sqrt{\rho(t)}$ has a bounded l_2-gain, neglecting $\gamma(P)$.

LEMMA 6.6.

$$\|w_{\tilde{\theta}}\|_t \leqq \gamma(H_{\tilde{\theta}})\|y\|_t+\gamma(D^{-1})\gamma(P)\beta_1(\|\phi\|_{t-1}+\alpha_1)+\gamma(D^{-1})\alpha_3\|w\|_t+\gamma(D^{-1})\alpha_2 M\sqrt{t}.$$

Proof.

$$\tilde{\theta}^T T(q)y(t) = \tilde{\theta}^T[B(q)(\phi(t-1)-W(q^{-1})y^m(t+d-1))+C(q)U(q^{-1})w(t)]^T,$$

where

$$\begin{aligned} U(q^{-1}) &:= (1, \cdots, q^{-n_S}, 0, \cdots, 0, 0, \cdots, 0)^T, \\ W(q^{-1}) &:= (0, \cdots, 0, 0, \cdots, 0, q^{-1}, \cdots, q^{-n_C})^T. \end{aligned}$$

From (3.1), (6.6), and the equality above, we have

$$D(q)w_{\tilde\theta}(t)=[D(q)-T^T(q)\tilde\theta]y(t)+P(q^{-1})\tilde\theta^T\phi(t-1)$$
$$-\tilde\theta^TB(q)W(q^{-1})y^m(t+d-1)+\tilde\theta^TC(q)U(q^{-1})w(t).$$

Using (6.3), we get

$$w_{\tilde\theta}(t)=H_{\tilde\theta}(q)y(t)+D^{-1}(q)P(q^{-1})\tilde\theta^T\phi(t-1)+D^{-1}(q)\tilde\theta^TC(q)U(q^{-1})w(t)$$
$$-D^{-1}(q)\tilde\theta^TB(q)W(q^{-1})y^m(t+d-1).$$

Hence

$$\|w_{\tilde\theta}\|_t\leqq\gamma(H_{\tilde\theta})\|y\|_t+\gamma(D^{-1})\gamma(P)R(\|\phi\|_{t-1}+\alpha_1)$$
$$+\gamma(D^{-1})R\gamma(B)\sqrt{n_C}M\sqrt{t}+\gamma(D^{-1})R\gamma(C)\sqrt{1+n_S}\|w\|_t.$$

If we choose $\alpha_2:=R\gamma(B)\sqrt{n_C}$, $\beta_1:=R$, and $\alpha_3:=R\gamma(C)\sqrt{1+n_S}$, the result follows. □

If we neglect the last three terms, this lemma and Lemma 6.5 tell us that the gain of operator $e(t)\to w_{\tilde\theta}(t)$ is $\gamma(H_{\tilde\theta})$.

LEMMA 6.7.

$$\|y^m\|_t\leqq M\sqrt{t}\leqq\alpha_4\rho^{1/2}(t)\quad\textit{where }\alpha_4:=M/\sqrt{\rho}.$$

Proof. From the fact that $|y^m(t)|\leqq M$ and $T\rho\leqq\rho(T)$, the result follows. □

Our result on the robustness of the adaptive controller with respect to the graph topology is given by the following theorem.

THEOREM 6.8. *Under assumptions* (A6.i)-(A6.iv), *the adaptive controller in feedback with the system* Π, *yields mean-square bounded inputs and outputs.*

Proof. As observed at the end of § 3, and due to the lemmas above, we can use a small gain argument.

Because $\rho(T)/T\geqq 1/T\sum_{t=1}^{T-d}y^2(t)$, if we can prove that there exists N so that

$$N>\frac{\rho(T)}{T},\tag{6.10}$$

then $\limsup_{T\to\infty}1/T\sum_{t=1}^{T}y^2(t)<\infty$. Similarly, we will have $\limsup_{T\to\infty}\cdot 1/T\sum_{t=1}^{T}u^1(t)<\infty$.

We now prove (6.10).

From Lemma 3.3(v), we get $\|e\|_T^2\leqq(1+\lambda_1)\sum_{t=1}^{T}(V_{\tilde\theta}(t-d)-V_{\tilde\theta}(t))\rho(t)+(1+\lambda_1)\|w_{\tilde\theta}\|_T^2$. Let

$$\Delta^2(T):=\max\left(0,\sum_{t=1}^{T}(V_{\tilde\theta}(t-d)-V_{\tilde\theta}(t))\frac{\rho(t)}{\rho(T)}\right)\quad\text{and}\quad\gamma_4:=(1+\lambda_1)^{1/2};\tag{6.11}$$

then

$$\|e\|_T\leqq\gamma_4\Delta(T)\rho^{1/2}(T)+\gamma_4\|w_{\tilde\theta}\|_T.\tag{6.12}$$

From Lemma 6.6 and Lemma 6.2(ii), (iii), we have

$$\|w_{\tilde\theta}\|_t\leqq\gamma(H_{\tilde\theta})\|y\|_t+\gamma(D^{-1})\gamma(P)\beta_1(\rho^{1/2}(t-1+d)+\alpha_1)$$
$$+\gamma(D^{-1})\alpha_3V\sqrt{t}+\gamma(D^{-1})\alpha_2M\sqrt{t}.\tag{6.13}$$

Substituting (6.13) into (6.12), with Lemmas 6.5 and 6.7 we obtain

$$\|e\|_T\leqq[\gamma_4\Delta(T)+\gamma_4\gamma(D^{-1})\gamma(P)\beta_1\gamma_3^{d-1}]\rho^{1/2}(T)+\gamma_4\gamma(D^{-1})\gamma(P)\beta_1\alpha_1$$
$$+\gamma_4\gamma(H_{\tilde\theta})\|e\|_T+[\gamma_4\gamma(H_{\tilde\theta})M+\gamma_4\gamma(D^{-1})(\alpha_3V+\alpha_2M)]\sqrt{T}.$$

Choose $\beta_2 := \beta_1\gamma_4$, $V_2 := \gamma_4 M$, $\delta_1 := \gamma_4\alpha_3$, $\beta_3 := \gamma_4\beta_1\alpha_1$, $\alpha_5 := \gamma_4\alpha_2 M$; then

(6.14)
$$(1-\gamma_4\gamma(H_{\tilde{\theta}}))\frac{\|e\|_T}{\sqrt{T}} \leqq (\gamma_4\Delta(T)+\beta_2\gamma(D^{-1})\gamma(P)\gamma_3^{d-1})\frac{\rho^{1/2}(T)}{\sqrt{T}} + V_2\gamma(H_{\tilde{\theta}})+\delta_1\gamma(D^{-1})V + \alpha_5\gamma(D^{-1})+\frac{\gamma(D^{-1})\gamma(P)\beta_3}{\sqrt{T}}.$$

From Lemmas 6.4 and 6.2 we have

$$\|u\|_{T-d} \leqq \gamma(D^{-1})[\gamma(A)\|y\|_T+\gamma(C)V\sqrt{T}+2\gamma(P)\rho^{1/2}(T+d-1)].$$

Using this inequality and Lemma 6.2, we have

$$\rho^{1/2}(T) \leqq \gamma_1\gamma(D^{-1})\gamma(A)\|y\|_T+\gamma_1\gamma(D^{-1})\gamma(C)\sqrt{T}V + 2\gamma_1\gamma(D^{-1})\gamma(P)\rho^{1/2}(T)\gamma_3^{d-1}+\gamma_2\|y\|_T+(\alpha_1+\sqrt{\rho})\sqrt{T-d}+\sqrt{V_1}.$$

Choose $\beta_4 = 2\gamma_1$, $\delta_2 = \gamma_1\gamma(C)$, $V_3 = \sqrt{\rho}$, and $\gamma_5 = \gamma_1\gamma(A)$; then

(6.15)
$$(1-\gamma(D^{-1})\gamma(P)\gamma_3^{d-1}\beta_4)\frac{\rho^{1/2}(T)}{\sqrt{T}} \leqq (\gamma(D^{-1})\gamma_5+\gamma_2)\frac{\|e\|_T}{\sqrt{T}}+(\gamma(D^{-1})\delta_2 V+\alpha_1+V_3+\gamma(D^{-1})\gamma_6+\gamma_7)+\frac{\sqrt{V_1}}{\sqrt{T}}$$

where we let $\gamma_6 := \gamma_5 M$, $\gamma_7 := \gamma_2 M$.

From assumption (A6.iii), there exists $\varepsilon > 0$, so that $1-\gamma(H_{\tilde{\theta}})\gamma_4 = \varepsilon$. Substituting (6.14) into (6.15), we get

(6.16)
$$\left(1-\gamma(D^{-1})\gamma(P)\beta_4\gamma_3^{d-1}-\frac{(\gamma(D^{-1})\gamma_5+\gamma_2)}{\varepsilon}(\gamma_4\Delta(T)+\beta_2\gamma_3^{d-1}\gamma(D^{-1})\gamma(P))\right)\frac{\rho^{1/2}(T)}{\sqrt{T}} \leqq M_1\frac{1}{\sqrt{T}}+M_0$$

where

$$M_0 := \gamma(D^{-1})\delta_2 V+\alpha_1+V_3+\gamma(D^{-1})\gamma_6+\gamma_7 + \frac{(\gamma(D^{-1})\gamma_5+\gamma_2)}{\varepsilon}[V_2\gamma(H_{\tilde{\theta}})+\alpha_5\gamma(D^{-1})+\gamma(D^{-1})\delta_1 V],$$

$$M_1 := \sqrt{V_1}+\frac{(\gamma(D^{-1})\gamma_5+\gamma_2)}{\varepsilon}\gamma(D^{-1})\gamma(P)\beta_3.$$

From assumption (A6.iv), we know that $\gamma(P)\gamma(D^{-1})(\beta_4+\beta_2((\gamma(D^{-1})\gamma_5+\gamma_2)/\varepsilon))\gamma_3^{d-1} < 1$. For convenience, define $\eta := 1-\gamma(D^{-1})\gamma(P)(\beta_4+\beta_2((\gamma(D^{-1})\gamma_5+\gamma_2)/\varepsilon))\gamma_3^{d-1} > 0$. Choose some fixed δ such that

(6.17)
$$0 < \delta < \frac{\eta\varepsilon}{\gamma_4(\gamma_2+\gamma_5\gamma(D^{-1}))}.$$

Case 1. For each time T such that $\Delta(T) \leqq \delta$, (6.16) can be rewritten as

$$\left(\eta-\frac{\gamma_4(\gamma_2+\gamma_5\gamma(D^{-1}))}{\varepsilon}\Delta(T)\right)\frac{\rho^{1/2}(T)}{\sqrt{T}} \leqq M_0+\frac{M_1}{\sqrt{T}}.$$

Let

$$M_2 := \eta-\frac{\gamma_4(\gamma_2+\gamma_5\gamma(D^{-1}))}{\varepsilon}\delta > 0 \quad \text{(from (6.17))};$$

then

$$\frac{\rho^{1/2}(T)}{\sqrt{T}} \leqq \frac{1}{M_2}\left(M_1 \frac{1}{\sqrt{T}} + M_0\right).$$

Hence $\rho(T)/T$ is bounded almost surely.

Case 2. Consider some time interval, say (T_0, T_1), such that

$$\begin{aligned} &\frac{\rho^{1/2}(T_0)}{\sqrt{T_0}} \leqq \left(\frac{M_1}{\sqrt{T_0}} + M_0\right)\frac{1}{M_2}, \\ &\frac{\rho^{1/2}(T)}{\sqrt{T}} > \left(\frac{M_1}{\sqrt{T}} + M_0\right)\frac{1}{M_2} \\ &\qquad\qquad \text{for every } T \in (T_0, T_1) \text{ (where } T_1 - T_0 \text{ may be infinite)}, \\ &\frac{\rho^{1/2}(T_1)}{\sqrt{T_1}} \leqq \left(\frac{M_1}{\sqrt{T_1}} + M_0\right)\frac{1}{M_2}. \end{aligned} \tag{6.18}$$

On such intervals, we necessarily have $\Delta(T) > \delta$ for every $T \in (T_0, T_1)$. From (6.11), $\Delta(T) > \delta$ yields

$$\sum_{t=1}^{T} (V_{\tilde{\theta}}(t-d) - V_{\tilde{\theta}}(t)) \frac{\rho(t)}{\rho(T)} > \delta^2. \tag{6.19}$$

Define $W_{\tilde{\theta}}(T) = \sum_{t=T-d+1}^{T} V_{\tilde{\theta}}(t)$, $T \geqq 0$. Note that $W_{\tilde{\theta}}(T) - W_{\tilde{\theta}}(T-1) = V_{\tilde{\theta}}(T) - V_{\tilde{\theta}}(T-d)$ for $T \geqq 1$. Because $0 \leqq V_{\tilde{\theta}}(t) \leqq V_4$, We have

$$0 \leqq W_{\tilde{\theta}}(T) \leqq dV_4. \tag{6.20}$$

From (6.19), we know that

$$\sum_{t=1}^{T} (W_{\tilde{\theta}}(t-1) - W_{\tilde{\theta}}(t)) \frac{\rho(t)}{\rho(T)} > \delta^2. \tag{6.21}$$

We define

$$W_{\tilde{\theta}}^{av}(T) = \left(\sum_{t=1}^{T} W_{\tilde{\theta}T} \frac{\rho(t+1) - \rho(t)}{\rho(T+1)}\right) + \frac{W_{\tilde{\theta}}(0)\rho(1)}{\rho(T+1)}.$$

Note the following:

(i) From (6.20), $0 \leqq W_{\tilde{\theta}}^{av}(T) \leqq dV_4$;

(ii) From (6.21), $\sum_{t=1}^{T} (W_{\tilde{\theta}}(t-1) - W_{\tilde{\theta}}(t))(\rho(t)/\rho(T)) > \delta^2$;

(iii) $$W_{\tilde{\theta}}^{av}(T) = W_{\tilde{\theta}}^{av}(T-1)\frac{\rho(T)}{\rho(T+1)} + \left(1 - \frac{\rho(T)}{\rho(T+1)}\right) W_{\tilde{\theta}}(T);$$

(iv) From (ii) and (iii),

$$W_{\tilde{\theta}}^{av}(T-1) - W_{\tilde{\theta}}^{av}(T) > \delta^2 \frac{\rho(T+1) - \rho(T)}{\rho(T+1)} \geqq \left(\frac{\delta}{\gamma_3}\right)^2 \frac{\rho(T+1) - \rho(T)}{\rho(T)}, \qquad T \in (T_0, T_1).$$

hence for any $T \in (T_0, T_1)$, we have $\sum_{t=T_0+1}^{T} (\rho(t+1) - \rho(t))/\rho(t) \leqq (\gamma_3/\delta)^2 V_4 d$. Since $\rho(t)$ is increasing,

$$\log \frac{\rho(T)}{\rho(T_0+1)} \leqq \left(\frac{\gamma_3}{\delta}\right)^2 V_4 d \quad \text{or} \quad \frac{\rho^{1/2}(T)}{\rho^{1/2}(T_0+1)} \leqq \exp\left(\frac{1}{2}\left(\frac{\gamma_3}{\delta}\right)^2 V_4 d\right).$$

From Lemma 6.3 and (6.18), we get

$$\frac{\rho^{1/2}(T)}{\sqrt{T}} \leqq \frac{\gamma_3}{M_2}\left(M_1 \frac{1}{\sqrt{T_0}} + M_0\right) \exp\left(\frac{1}{2}\left(\frac{\gamma_3}{\delta}\right)^2 V_4 d\right).$$

Cases 1 and 2 tell us $\rho^{1/2}(T)/\sqrt{T}$ is bounded. □

7. Stability with nonvanishing adaptive gains. In the previous sections the gain of the parameter estimates, or equivalently the stepsize of the adaptation algorithm, has been allowed to converge asymptotically to zero. Indeed this is necessary if we asymptotically want to achieve optimal tracking. However, this vanishing gain also causes the adaptive controller to have asymptotically *diminishing* ability to adjust to *system changes.* Hence, in practice, adaptation gains are frequently prevented from going to zero. Therefore in this section we address the nonvanishing gain case of our adaptive controller.

We choose μ such that $0 < \mu < 1$. Let Γ' be the set of proper rational functions $F(q)$ whose poles are all in the open disk of radius μ. Γ' is equipped with the norm

$$\gamma(F) = \sup_{|q|=\mu} |F(q)|, \qquad F \in \Gamma'.$$

For a sequence $x(t)$, we define its $l_2(\mu)$-norm as $\|x\|_T^2 := \sum_{t=1}^{T} \mu^{-2t} x^2(t)$. Note that if $F(q) \in \Gamma'$ and $z(t) = F(q)x(t)$, then $\|z\|_T \leqq \gamma(F)\|x\|_T$.

Let us consider system Π, which can be described as follows:

(A7.i) We suppose that the true system satisfies

$$A(q)y(t) = B(q)u(t-1) + C(q)w(t), \qquad t \geqq 1 \tag{7.1}$$

where A, B, $C \in \Gamma'$, A, B are coprime, B/A is a proper rational function, and $A(\infty) = 1 = C(\infty)$. Regarding the noise $w(t)$ we will merely assume that it is bounded, $|w(t+1)| \leqq V$, where V is a *deterministic* finite positive number.

As before we will also assume that $|y^m(t)| \leqq M$ for all $t > 0$ and

$$y^m(t) = u(t) = y(t) = w(t) = 0 \quad \text{for all } t \leqq 0.$$

Because $B(q)$ is an analytic function outside the disk of radius μ, we can write a Laurent series $B(q) = \sum_{i=0}^{\infty} h_i q^{-i}$ and state that if $d \geqq 2$, then $P(d^{-1}) = \sum_{i=0}^{d-2} h_i q^{-i}$; otherwise it equals zero and $D(q) = \sum_{i=0}^{\infty} h_{i+d-1} q^{-i}$. It is easy to see that $B(q) = P(q^{-1}) + q^{1-d} D(q)$. Then we assume the following:

(A7.ii) $D(q)$ is an invertible element of Γ'.

As before, we define

$$T(q) := (A(q), \cdots, q^{-n_S} A(q), q^{-1} B(q), \cdots, q^{-(n_R+1)} B(q), 0, \cdots, 0)^T.$$

For any θ, we define $H_\theta(q) := 1 - D^{-1}(q) T^T(q)\theta$ and choose $\tilde{\theta} \in \Theta$ so that $\gamma(H_{\tilde{\theta}}) \leqq \gamma(H_\theta)$, for all $\theta \in \Theta$. We also make the following two assumptions:

(A7.iii) $\gamma(H_{\tilde{\theta}}) < \gamma_h$ where $\gamma_h := 1/\gamma_4$;

(A7.iv) $\gamma(P) < (\gamma_h - \gamma(H_{\tilde{\theta}}))/(\gamma(D^{-1})(k_4 \gamma(D^{-1})\gamma(A) + k_5 + k_6(\gamma_h - \gamma(H_{\tilde{\theta}})))\gamma_{10}^{d-1})$ where k_4, k_5, and k_6 are strictly positive constants given in Table 1 in the Appendix (as is γ_{10} also).

Because the mapping $F(q) \to F(\mu q)$ is an isomorphism on the field of rational functions, all the properties of [26] can be used here. In particular we obtain a topology that is the weakest one, such that feedback μ-exponential stability is robust.

THEOREM 7.1. *The set of* (A, B, C) *satisfying assumptions* (A7.ii)-(A7.iv) *is open.*

Proof. The proof is the same as in Theorem 6.1 except that we need to prove that the mapping $(A, B, C) \to P$ is continuous. Note first that $\gamma(P) \leqq \sum_{i=0}^{d-2} |h_i| \mu^{-i}$ and $\sum_{i=0}^{d-2} |h_i| \mu^{-i} \leqq \sqrt{d-1} (\sum_{i=0}^{\infty} h_i^2 \mu^{-2i})^{1/2}$. Now, using Parseval's theorem, we have $\sum_{i=0}^{\infty} h_i^2 \mu^{-2i} \leqq \sup_{|q| \geqq \mu} |B(q)|^2$. Hence $\gamma(P) \leqq \sqrt{d-1}\gamma(B)$ proves that the mapping $(A, B, C) \to P$ is continuous. □

Now we define a *new* normalization sequence:

$$\rho(t) = \mu^2 \rho(t-1) + \max(\rho, \|\phi(t-d)\|^2), \qquad t \geqq 1 \tag{7.2}$$

where $\rho(t) = 0$ if $t \leqq 0$ and $0 < \rho < \infty$.

It is important to note that in going from (2.1), where we had simply $\mu^2 = 1$, to (7.2), we have made our assumptions more restrictive. This can be seen by comparing Theorems 6.1 and 7.1, In the latter we need μ-exponential stability, whereas in the former mere exponential stability is sufficient. In particular, this means that in the latter case we cannot neglect a pole-zero pair that nearly cancels and that corresponds to an eigenvalue larger than μ in modulus. We can also note that for the first example of § 6, we now obtain the restriction $|a_2| < \mu/\sqrt{1+\lambda_1}$.

The following lemmas are essentially similar to those in § 6, and so we abbreviate the proofs.

LEMMA 7.2.

(i) $\mu^{-T}|w(T)| \leqq \|w\|_T \leqq \delta_3 \mu^{-T} V$;

(ii) $\mu^{-2t}\rho(t) \geqq \mu^{-2(t-1)}\rho(t-1)$;

(iii) $\frac{1}{2}(\|u\|_T + \|y\|_T) \leqq \|\phi\|_T \leqq \gamma_9 \|u\|_T + \gamma_{11}\|y\|_T + \alpha_6 \mu^{-T}$;

(iv) $\|\phi\|_T \leqq \gamma_8 \mu^{-(T+d)} \rho^{1/2}(T+d) \leqq \|\phi\|_T + V_6 \mu^{-T} + V_5$.

Proof. (i) The proof follows from the definition of the norm $\|w\|_T$ as $\|w\|_T = \sum_{t=1}^{T} \mu^{-2t} w^2(t)$ and from (A7.i), which assumes $|w(t+1)| \leqq V$.

Inequality (ii) follows from (7.2).

(iii) $\|\phi\|_T^2 \leqq (1+n_S)\mu^{-2n_S}\|u\|_T^2 + (1+n_R)\mu^{-2n_R}\|y\|_T^2 + n_C M^2(\mu^{-2T}/(1-\mu^2))$. Now choose $\gamma_9 = \sqrt{1+n_S}\,\mu^{-n_S}$, $\gamma_{11} = \sqrt{1+n_R}\,\mu^{-n_R}$, and $\alpha_6 = M(\sqrt{n_C}/\sqrt{1-\mu^2})$ and the result follows.

(iv) $\mu^{-2(T+d)}\rho(T+d) \geqq \sum_{t=1}^{T+d} \mu^{-2t}\|\phi(t-d)\|^2 \geqq \mu^{-2d}\|\phi\|_T^2$.

When we choose $\gamma_8 = \mu^d$, the left inequality in (iv) follows. On the other hand, when we use

$$\sum_{t=1}^{T+d} \mu^{-2t}(\|\phi(t-d)\|^2 + \rho) \geqq \mu^{-2(T+d)}\rho(T+d) - \rho(0),$$

$$\mu^{-2d}\left(\frac{1-\mu^{2d}}{1-\mu^2}\alpha_1^2 + \|\phi\|_T^2\right) \geqq \sum_{t=1}^{T+d} \mu^{-2t}\|\phi(t-d)\|^2 \quad (\text{where } \alpha_1 := M\sqrt{n_C}),$$

$$\sum_{t=1}^{T+d} \mu^{-2t}\rho = \frac{\mu^{-2}(1-\mu^{-2(T+d)})}{1-\mu^2} p,$$

the right-hand side inequality in (iv) holds if we choose

$$V_5 = \alpha_1 \frac{\sqrt{1-\mu^{2d}}}{\sqrt{1-\mu^2}} \quad \text{and} \quad V_6 = \frac{\sqrt{\rho}}{\sqrt{1-\mu^2}}.$$

□

LEMMA 7.3. $\rho(t-1) \leqq \gamma_{10}^2 \rho(t)$.

Proof. If we use $|y(t+1)| \leqq |w(t+1)| + \mu t\gamma(B)\|u\|_t + \mu^{t+1}\gamma(A-1)\|y\|_t + \mu^{t+1}\gamma(C-1)\|w\|_t$ instead of (6.4), and

$$y^2(t-d+1) \leqq [|w(t-d+1)| + (\gamma(B)+\mu\gamma(A-1))\mu^{t-d} \cdot (\|u\|_{t-d} + \|y\|_{t-d}) + \mu^{t-d+1}\gamma(C-1)\|w\|_{t-d}]^2$$

$$\leqq \left[2(\gamma(B)+\mu\gamma(A-1))\mu^{-d}\gamma_8 + (1+\gamma(C-1)\mu)\delta_3 \frac{V}{\sqrt{\rho}}\right]^2 \rho(t)$$

instead of (6.9), then we get the desired result. □

LEMMA 7.4.

$$\|u\|_{T-d} \leqq \mu^d \gamma(D^{-1})(\gamma(A)\|y\|_T + \gamma(C)\|w\|_T + \mu^{-1}\gamma(P)\|u\|_{T-1}).$$

Proof. Because $D(q)u(t-d) = A(q)y(t) - C(q)w(t) - P(q^{-1})u(t-1)$, we have $\|u\|_{T-d} \leqq \mu^d\gamma(D^{-1})(\|Ay\|_T + \|Cw\|_T + \mu^{-1}\|Pu\|_{T-1})$. □

The next result is immediate.

LEMMA 7.5.

(i) $\|y\|_T \leqq \|y^m\|_T + \|e\|_T$;

(ii) $\|y^m\|_T \leqq M(\mu^{-2T}(1-\mu^{2T})/(1-\mu^2))^{1/2} \leqq \alpha_8\mu^{-T}$.

LEMMA 7.6.

$$\|w_{\tilde{\theta}}\|_T \leqq \gamma(H_{\tilde{\theta}})\|y\|_T + \gamma(D^{-1})\gamma(P)\beta_6(\|\phi\|_{T-1} + \alpha_1) + \gamma(D^{-1})\alpha_9\mu^{-T} + \gamma(D^{-1})\alpha_{10}\|w\|_T.$$

Proof. The proof is similar to that of Lemma 6.6. □

THEOREM 7.7. *Under assumptions* (A7.i)-(A7.iv), *the adaptive controller in feedback with the system* Π, *yields bounded inputs and outputs almost surely.*

Proof. From Lemma 7.2(iv) and (iii), we have $\frac{1}{2}\mu^{-T}y(T) \leqq \mu^{-T}\rho^{1/2}(T+d)$. If we can now prove that there exists $N < \infty$ almost surely so that

$$\rho^{1/2}(T) < N \quad \text{for } T > 0, \tag{7.3}$$

then clearly $y(T) \leqq 2\rho^{1/2}(T+d) < 2N$, so we will have shown that $y(T)$ is bounded almost surely. A similar situation holds regarding $u(T)$ also.

So we only need to prove (7.3). From Lemma 3.3(v), we have

$$\|e\|_T^2 \leqq (1+\lambda_1)\mu^{-2T}\rho(T)\Delta^2(T) + (1+\lambda_1)\|w_{\tilde{\theta}}\|_T^2$$

where

$$\Delta^2(T) := \max\left(0, \sum_{t=1}^{T} (V_{\tilde{\theta}}(t-d) - V_{\tilde{\theta}}(t))\mu^{2(T-t)}\frac{\rho(t)}{\rho(T)}\right). \tag{7.4}$$

Thus

$$\|e\|_T \leqq \gamma_4\mu^{-T}\rho^{1/2}(T)\Delta(T) + \gamma_4\|w_{\tilde{\theta}}\|_T \quad \text{where } \gamma_4 = \sqrt{1+\lambda_1}. \tag{7.5}$$

From Lemmas 7.6 and 7.2, we have

$$\|w_{\tilde{\theta}}\|_T \leqq \gamma(H_{\tilde{\theta}})\|y\|_T + \gamma(D^{-1})\gamma(P)\beta_7\mu^{-T}\rho^{1/2}(T)\gamma_{10}^{d+1} + \gamma(D^{-1})\gamma(P)\delta_4 + \gamma(D^{-1})(\alpha_9 + \alpha_{10}\delta_3 V)\mu^{-T} \tag{7.6}$$

where $\beta_7 := \beta_6\gamma_8\mu^{1-d}$ and $\delta_4 := \beta_6\alpha_1$. Substituting (7.6) into (7.5), we obtain

$$\begin{aligned}\|e\|_T \leqq{}& \gamma_4\mu^{-T}\rho^{1/2}(T)\Delta(T) + \gamma_4\gamma(H_{\tilde\theta})\|y\|_T + \gamma(D^{-1})\gamma(P)\beta_8\mu^{-T}\rho^{1/2}(T)\gamma_{10}^{d-1} \\ &+ \gamma(D^{-1})\gamma(P)\delta_5 + \gamma(D^{-1})(\alpha_{11}+\alpha_{12}V)\mu^{-T}\end{aligned} \tag{7.7}$$

where $\beta_8 := \gamma_4\beta_7$, $\delta_5 := \gamma_4\delta_4$, $\alpha_{11} := \alpha_9\gamma_4$, and $\alpha_{12} := \alpha_{10}\delta_3\gamma_4$.

From Lemma 7.5, we have $\|y\|_T \leqq \|y^m\|_T + \|e\|_T \leqq \alpha_8\mu^{-T} + \|e\|_T$. Using this, we rewrite (7.7) as

$$\begin{aligned}(1-\gamma_4\gamma(H_{\tilde\theta}))\|e\|_T \leqq{}& [\gamma_4\Delta(T) + \gamma(D^{-1})\gamma(P)\beta_8\gamma_{10}^{d-1}]\mu^{-T}\rho^{1/2}(T) + \gamma(D^{-1})\gamma(P)\delta_5 \\ &+ [\gamma_4\gamma(H_{\tilde\theta})\alpha_8 + \gamma(D^{-1})(\alpha_{11}+\alpha_{12}V)]\mu^{-T}.\end{aligned} \tag{7.8}$$

From Lemmas 7.4, 7.2, and 7.3, we have

$$\|u\|_{T-d} \leqq \gamma(D^{-1})[\gamma_{15}\|y\|_T + \gamma(P)\delta_7\mu^{-T}\rho^{1/2}(T)\gamma_{10}^{d-1} + \delta_8\mu^{-T}V]$$

where $\gamma_{15} := \mu^d\gamma(A)$, $\delta_7 := 2\gamma_8\mu$, and $\delta_8 := \gamma(C)\mu^d\delta_3$.

From Lemma 7.2(iv), (iii), and this last inequality, we have

$$\begin{aligned}\mu^{-T}\rho^{1/2}(T) &\leqq \frac{\gamma_9}{\gamma_8}\|u\|_{T-d} + \frac{\gamma_{11}}{\gamma_8}\|y\|_{T-d} + \frac{\alpha_6}{\gamma_8}\mu^{-(T-d)} + \frac{V_5}{\gamma_8} + \frac{V_6}{\gamma_8}\mu^{-(T-d)} \\ &\leqq \left(\frac{\gamma_9}{\gamma_8}\gamma(D^{-1})\gamma_{15} + \frac{\gamma_{11}}{\gamma_8}\right)\|y\|_T + \frac{\gamma_9}{\gamma_8}\gamma(D^{-1})\gamma(P)\delta_7\mu^{-T}\rho^{1/2}(T)\gamma_{10}^{d-1} \\ &\quad + \frac{\gamma_9}{\gamma_8}\gamma(D^{-1})\delta_8\mu^{-T}V + \frac{1}{\gamma_8}\alpha_6\mu^d\mu^{-T} + \frac{V_5}{\gamma_8} + \frac{V_6}{\gamma_8}\mu^{-T}\mu^d.\end{aligned}$$

If we choose

$$\begin{aligned}&\delta_9 := \frac{\gamma_9}{\gamma_8}\delta_7, \quad \gamma_{14} := \frac{\gamma_9}{\gamma_8}\gamma_{15}, \quad \gamma_{13} := \frac{\gamma_{11}}{\gamma_8}, \quad \delta_6 := \frac{\gamma_9}{\gamma_8}\delta_8, \\ &\alpha_7 := \frac{1}{\gamma_8}\alpha_6\mu^d, \quad V_7 := \frac{V_5}{\gamma_8}, \quad V_8 := \frac{V_6}{\gamma_8}\mu^d,\end{aligned}$$

then the inequality above becomes

$$\begin{aligned}&(1-\gamma(D^{-1})\delta_9\gamma(P)\gamma_{10}^{d-1})\mu^{-T}\rho^{1/2}(T) \\ &\quad \leqq [\gamma(D^{-1})\gamma_{14}\alpha_8 + \gamma_{13}\alpha_8 + \gamma(D^{-1})\delta_6 V + \alpha_7 + V_8]\mu^{-T} \\ &\quad\quad + [\gamma(D^{-1})\gamma_{14} + \gamma_{13}]\|e\|_T + V_7.\end{aligned} \tag{7.9}$$

From assumption (A7.iii) there exists $\varepsilon > 0$ such that $1-\gamma(H_{\tilde\theta})\gamma_4 = \varepsilon$. Substituting (7.8) into (7.9), we have

$$\begin{aligned}&\left[1-\gamma(D^{-1})\delta_9\gamma(P)\gamma_{10}^{d-1} - \frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}\right. \\ &\qquad \left.\cdot\,(\gamma_4\Delta(T) + \gamma(D^{-1})\gamma(P)\beta_8\gamma_{10}^{d-1})\right]\rho^{1/2}(T) \leqq M_0 + \mu^T M_1\end{aligned} \tag{7.10}$$

where

$$\begin{aligned}M_0 :={}& \frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}[\gamma_4\gamma(H_{\tilde\theta})\alpha_8 + \gamma(D^{-1})(\alpha_{11}+\alpha_{12}V)] + \gamma(D^{-1})\gamma_{14}\alpha_8 + \gamma_{13}\alpha_8 \\ &+ \gamma(D^{-1})\delta_6 V + \alpha_7 + V_8, \\ M_1 :={}& \frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}\gamma(D^{-1})\gamma(P)\delta_5 + V_7.\end{aligned}$$

Now, assumption (A7.iv) implies

$$\gamma(P)\gamma(D^{-1})\left(\delta_9+\beta_8\frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}\right)\gamma_{10}^{d-1}<1.$$

Hence, there exists $\eta>0$ so that

$$\eta:=1-\gamma(D^{-1})\gamma(P)\left(\delta_9+\frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}\beta_8\right)\gamma_{10}^{d-1}.$$

Now we choose and fix a $\delta>0$ that satisfies $0<\delta<(\eta\varepsilon/(\gamma_4[\gamma(D^{-1})\gamma_{14}+\gamma_{13}]))$ and examine two cases.

Case 1. If for each time T, $\Delta(T)\leqq\delta$, then from (7.10),

$$\left(\eta-\frac{\gamma(D^{-1})\gamma_{14}+\gamma_{13}}{\varepsilon}\gamma_4\Delta T\right)\rho^{1/2}(T)\leqq M_0+\mu M_1.$$

Hence

$$\rho^{1/2}(T)\leqq\frac{\varepsilon(M_0+\mu M_1)}{\eta\varepsilon-[\gamma(D^{-1})\gamma_{14}+\gamma_{13}]\delta\gamma_4}=:N_1$$

and so $\rho(T)$ is bounded.

Case 2. Suppose, however, that there is a certain time interval (T_0, T_1) such that

$$\rho^{1/2}(T_0)\leqq N_1,$$

$$\rho^{1/2}(T)>N_1 \quad \text{for } T\in(T_0, T_1) \quad \text{(where } T_1-T_0 \text{ may be infinite)}, \tag{7.11}$$

$$\rho^{1/2}(T_1)\leqq N_1.$$

On such an interval, we must have $\Delta(T)>\delta$ for $T\in(T_0, T_1)$. From (7.4) we get

$$\sum_{t=1}^{T}(V_{\tilde{\theta}}(t-d)-V_{\tilde{\theta}}(t))\mu^{2(T-d)}\frac{\rho(t)}{\rho(T)}<\delta^2.$$

Let us define $W_{\tilde{\theta}}(T):=\sum_{t=T-d+1}^{T}V_{\tilde{\theta}}(t)$, for $T\geqq 0$, then

$$\sum_{t=1}^{T}(V_{\tilde{\theta}}(t-d)-V_{\tilde{\theta}}(t))\frac{\mu^{-2t}\rho(t)}{\mu^{-2T}\rho(T)}=\sum_{t=1}^{T}(W_{\tilde{\theta}}(t-1)-W_{\tilde{\theta}}(t))\frac{\mu^{-2t}\rho(t)}{\mu^{-2T}\rho(T)}>\delta^2.$$

Note that $0\leqq W_{\tilde{\theta}}(T)\leqq dV_4$, for $T\geqq 0$.

If we now define

$$W_{\tilde{\theta}}^{av}(T):=\sum_{t=1}^{T}W_{\tilde{\theta}}(t)\frac{\mu^{-2(t+1)}\rho(t+1)-\mu^{-2t}\rho(t)}{\mu^{-2(T+1)}\rho(T+1)}+\frac{W_{\tilde{\theta}}(0)\mu^{-2}\rho(1)}{\mu^{-2(T+1)}\rho(T+1)},$$

then

$$\sum_{t=1}^{T}(W_{\tilde{\theta}}(t-1)-W_{\tilde{\theta}}(t))\frac{\mu^{-2t}\rho(t)}{\mu^{-2T}\rho(T)}=W_{\tilde{\theta}}^{av}(T-1)-W_{\tilde{\theta}}(T)>\delta^2. \tag{7.12}$$

Note that because $W_{\hat{\theta}}(T) \leqq dV_4$, it follows that $W_{\hat{\theta}}^{av}(T) \leqq dV_4$. It is easy to see that

$$W_{\hat{\theta}}^{av}(T) = \frac{\mu^{-2T}\rho(T)}{\mu^{-2(T+1)}\rho(T+1)} W_{\hat{\theta}}^{av}(T-1) + \left(1 - \frac{\mu^{-2T}\rho(T)}{\mu^{-2(T+1)}\rho(T+1)}\right) W_{\hat{\theta}}(T).$$

Using (7.12) we have

$$\begin{aligned} & W_{\hat{\theta}}^{av}(T-1) - W_{\hat{\theta}}^{av}(T) \\ & > \left(1 - \frac{\mu^{-2T}\rho(T)}{\mu^{-2(T+1)}\rho(T+1)}\right)(\delta^2 + W_{\hat{\theta}}(T)) - \left(1 - \frac{\mu^{-2T}\rho(T)}{\mu^{-2(T+1)}\rho(T+1)}\right) W_{\hat{\theta}}(T) \\ & \geqq \left(\frac{\delta\mu}{\gamma_{10}}\right)^2 \left(\frac{\mu^{-2(T+1)}\rho(T+1) - \mu^{-2T}\rho(T)}{\mu^{-2T}\rho(T)}\right). \end{aligned}$$

Hence,

$$\begin{aligned} \sum_{t=T_0+1}^{T} \frac{\mu^{-2(t+1)}\rho(t+1) - \mu^{-2t}\rho(t)}{\mu^{-2t}\rho(t)} & \leqq \sum_{t=T_0+1}^{T} \left(\frac{\gamma_{10}}{\delta\mu}\right)^2 [W_{\hat{\theta}}^{av}(t-1) - W_{\hat{\theta}}^{av}(t)] \\ & \leqq \left(\frac{\gamma_{10}}{\delta\mu}\right)^2 dV_4 \quad \text{for } T \in (T_0, T_1). \end{aligned}$$

Because $\mu^{-2t}\rho(t)$ is increasing with respect to time t, from the last inequality we get

$$\frac{\mu^{-2T}\rho(T)}{\mu^{-2(T_0+1)}\rho(T_0+1)} < \exp\left(\left(\frac{\gamma_{10}}{\delta\mu}\right)^2 dV_4\right).$$

and so again $\rho(T)$ is bounded.

When we combine Cases 1 and 2, the theorem is proved. □

8. Conclusions. Here we have analyzed the twin issues of obtaining both *good performance* and *robustness* out of an adaptive controller for linear stochastic systems.

For minimum phase plants of known order, with a known compact set containing a stabilizing regulator, and for which we know the sign and a lower bound for the leading coefficient of the control polynomial, we have shown that our adaptive controller yields *mean-square bounded* inputs and outputs. If the noise additionally satisfies a positive real condition, then we have shown that the adaptive controller is asymptotically *optimal* in the sense of *minimizing* output error variance. We have also presented a *graph topological neighborhood* of an ideal plant, such that the system is mean-square stabilized even when that system is not ideal and when the statistical properties of the noise are violated. This holds true whether the adpative controller is used in a vanishing or a nonvanishing gain mode.

Several open problems remain. It is still not known whether the standard self-tuning regulator using a *least-squares* parameter estimate is mean-square stable, let alone optimal. Moreover, it is not known whether the unmodified adaptive controller possesses good robustness properties. The first question deals essentially with the loss of identifiability, and the consequent unboundedness of the condition number of the so-called "covariance matrix," when the parameter estimates converge. Unfortunately the second issue cannot really be resolved until the first issue is better understood.

Appendix.

TABLE 1

$\alpha_1 = \sqrt{n_C} M$
$\alpha_2 = R\gamma(B)\sqrt{n_C}$
$\alpha_3 = R\gamma(C)\sqrt{1+n_S}$
$\alpha_4 = M/\sqrt{\rho}$
$\alpha_5 = \alpha_2 M \gamma_4$
$\alpha_6 = (\sqrt{n_C}/\sqrt{1-\mu^2})M$
$\alpha_7 = (\alpha_6/\gamma_8)\mu^d$
$\alpha_8 = M/\sqrt{1-\mu^2}$
$\alpha_9 = R\gamma(B)\sqrt{n_C}\mu^{-n_C}\alpha_8$
$\alpha_{10} = R\gamma(C)\sqrt{1+n_S}\mu^{-n_S}$
$\alpha_{11} = \alpha_9\gamma_4$
$\alpha_{12} = \alpha_{10}\delta_3\gamma_4$
$\delta_1 = \gamma_4\alpha_3$
$\delta_2 = \gamma_1\gamma(C)$
$\delta_3 = 1/\sqrt{1-\mu^2}$
$\delta_4 = \beta_6\alpha_1$
$\delta_5 = \gamma_4\delta_4$
$\delta_6 = (\gamma_9/\gamma_8)\delta_8$
$\delta_7 = 2\gamma_8\mu$
$\delta_8 = \mu^d\delta_3\gamma(C)$
$\delta_9 = (\gamma_9/\gamma_8)\delta_7$
$k_3 = \beta_4$
$k_4 = \beta_8\gamma_{14}/\gamma_4$
$k_5 = \beta_8\gamma_{13}/\gamma_4$
$k_6 = \delta_9$
$k_7 = \gamma_4$
$k_8 = \gamma_4\beta_6\gamma_{10}^d$
$k_9 = \gamma_4\beta_9(M/\sqrt{\rho})$
$k_{10} = \gamma_{10}\gamma_4(\delta_3/\sqrt{\rho})$
$\gamma_1 = \sqrt{1+n_S}$
$\gamma_2 = \sqrt{1+n_R}$
$\gamma_4 = \sqrt{1+\lambda_1}$

$\beta_1 = R$
$\beta_2 = \beta_1\gamma_4$
$\beta_3 = \gamma_4\beta_1\alpha_1$
$\beta_4 = 2\gamma_1$
$\beta_5 = \gamma_4\sqrt{dV_4}$
$\beta_6 = R/\mu$
$\beta_7 = \beta_6\gamma_8\mu^{1-d}$
$\beta_8 = \gamma_4\beta_7$
$\beta_9 = R\gamma(B)\sqrt{n_C}\mu^{-n_C}$
$V_1 = \rho(d)$
$V_2 = \gamma_4 M$
$V_3 = \sqrt{\rho}$
$V_4 = 1/\lambda_0(K + K(\lambda_1/\lambda_0))$
$V_5 = \alpha_1(\sqrt{1-\mu^{2d}}/\sqrt{1-\mu^2})$
$V_6 = \sqrt{\rho}/\sqrt{1-\mu^2}$
$V_7 = V_5/\gamma_8$
$V_8 = \mu^d(V_6/\gamma_8)$
$k_1 = \beta_2\gamma_1/\gamma_4$
$k_2 = \beta_2\gamma_2/\gamma_4$
$\gamma_5 = \gamma_1\gamma(A)$
$\gamma_6 = \gamma_5 M$
$\gamma_7 = \gamma_2 M$
$\gamma_8 = \mu^d$
$\gamma_9 = \mu^{-n_S}\sqrt{1+n_S}$
$\gamma_{11} = \mu^{-n_R}\sqrt{1+n_R}$
$\gamma_{12} = \mu\gamma_1(\gamma_9/\gamma_8)$
$\gamma_{13} = \gamma_{11}/\gamma_8$
$\gamma_{14} = \gamma_{15}(\gamma_9/\gamma_8)$
$\gamma_{15} = \mu^d\gamma(A)$
$\gamma_h = 1/\sqrt{1+\lambda_1}$

$$\gamma_3^2 = 2 + \frac{2}{\sigma_0^2}R^2 + \frac{M^2}{\rho} + 1 + \frac{2}{\rho\sigma_0^2}M^2(1+R^2) + \left(1+\frac{2}{\sigma_0^2}R^2\right)\left[\frac{V}{\sqrt{\rho}} + 2(\gamma(B)+\gamma(A-1)) + \frac{\gamma(C-1)}{\sqrt{\rho}}V\right]^2,$$

$$\gamma_{10}^2 = 1+\mu^2+\frac{2}{\sigma_0^2}R^2 + \frac{1}{\rho}\left(\rho + \frac{2}{\sigma_0^2}M^2(1+R^2) + M^2\right) + \left(1+\frac{2}{\sigma_0^2}R^2\right)\left[2(\gamma(B)+\gamma(A-1))\mu^{-d}\gamma_8 + (1+\gamma(C-1)\mu)\delta_3\frac{V}{\sqrt{\rho}}\right]^2$$

REFERENCES

[1] K. J. ASTRÖM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.

[2] K. J. ASTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 195–199.

[3] A. H. BECKER, P. R. KUMAR, AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm: geometry and convergence*, IEEE Trans. Atomat. Control, AC-30 (1985), pp. 330–338.

[4] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[5] B. EGARDT, *Stability of Adaptive Controllers*, Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York, 1979.

[6] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19 (1981), pp. 829–853.

[7] P. IOANNOU AND P. KOKOTOVIC, *An asymptotic error analysis of identifiers and adaptive observers and in the presence of parasitics*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 921-927.
[8] P. IOANNOU AND K. TSAKALIS, *Robust discrete-time adaptive control*, in Adaptive and Learning Systems: Theory and Applications, K. S. Narendra, ed., Plenum Press, New York, 1986.
[9] G. KREISSLEMEIER AND B. D. O. ANDERSON, *Robust model reference adaptive control*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 127-133.
[10] P. R. KUMAR, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329-380.
[11] P. R. KUMAR AND L. PRALY, *Self-tuning trackers*, SIAM J. Control Optim., 25 (1987), pp. 1053-1071.
[12] P. R. KUMAR AND P. VARAIYA, *Stochastic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
[13] T. L. LAI AND C. Z. WEI, *Least squares estimate in stochastic regression with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154-166.
[14] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551-575.
[15] ———, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Automat Control, AC-22 (1977), pp. 539-551.
[16] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.
[17] V. PETERKA, *On steady state minimum variance control strategy*, Kybernetika, 8 (1972), pp. 218-231.
[18] L. PRALY, *Stochastic adaptive controllers with and without positivity condition*, in Proc. 23rd IEEE Conference on Decision and Control, December 1984.
[19] ———, *Global stability of a direct adaptive control scheme is robust with respect to a graph topology*, in Adaptive and Learning Systems: Theory and Applications, K. S. Narendra, ed., Plenum Press, New York, 1986.
[20] ———, *Robust model reference adaptive controllers, Part* I: *Stability analysis*, in Proc. 23rd IEEE Conference on Decision and Control, December 1984.
[21] ———, *Robustness of model reference adaptive control*, in Proc. Yale Workshop on Applications of Adaptive Systems Theory, June 1983.
[22] C. ROHRS, L. VALAVANI, M. ATHANS, AND G. STEIN, *Analytical verification of undesirable properties of direct model*, in Proc. 20th IEEE Conference on Decision and Control, 1981, pp. 1272-1284.
[23] U. SHAKED AND P. R. KUMAR, *Minimum variance control of multivariable* ARMAX *systems*, SIAM J. Control Optim., 24 (1986), pp. 396-411.
[24] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 315-321.
[25] V. SOLO, *The convergence of* AML, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958-962.
[26] M. VIDYASAGAR, *The graph metric for unstable plants and robustness estimates for feedback stability*, IEEE Trans. Automat. Control, AC-29, (1984), pp. 403-418.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 267-288, March 1989

002

PROPERTIES OF THE RELAXED TRAJECTORIES OF EVOLUTION EQUATIONS AND OPTIMAL CONTROL*

NIKOLAOS S. PAPAGEORGIOU†

Abstract. This paper considers a class of nonconvex control problems in a Banach space, governed by a semilinear evolution equation and establishes the existence of admissible pairs. Then the introduction of measure-valued controls convexifies the system, and under general assumptions it is shown that the set of trajectories of the original system is dense (in an appropriate topology) in the set of relaxed trajectories. Some useful topological properties of the set of relaxed trajectories are also determined. Furthermore, some optimization problems are solved and in many cases the values of the original and relaxed problems are shown to be equal. Finally, another relaxation result is proved for a different class of systems, described by a general nonlinear evolution equation.

Key words. evolution inclusion, measurable multifunction selection, evolution operator, weak compactness

AMS(MOS) subject classification. 49A20

1. Introduction.

It is well known to researchers working in optimal control theory that to guarantee existence of optimal "state-control" pairs we need, among other things, a convexity hypothesis on a certain orientor field. The well-known property (Q) of Cesari is an example of such a hypothesis. It was first introduced by Cesari [12a] to prove the lower semicontinuity of the cost functional and since then it has become a popular tool among people working on variational and optimal control problems. This is very nicely exemplified in Berkovitz [7] and Cesari [13].

When this convexity hypothesis is no longer satisfied, to have optimal solutions we need to pass to a larger system, in which the orientor field (also known as the velocity field or tangent bundle) has been convexified. Such a convexification, on a control theoretic level, corresponds to the introduction of measure-valued controls. Those controls are called "relaxed controls" (see Warga [37], [38]) or "chattering controls" (see Gamkrelidze [22]) or "sliding regimes" (see Filippov [21]) or "generalized curves" (see Young [39]). Here we adopt the name "relaxed control," which is more widely used among mathematicians working in control theory. Having introduced this augmented system, it is natural to ask what the relation is between the set of relaxed trajectories and the set of original trajectories. More precisely, given the fact that under mild assumptions the relaxed problem has a solution, we would like to know whether every relaxed optimal trajectory can be approximated arbitrarily close by trajectories of the original system.

For finite-dimensional control systems this problem has already been studied in the literature. For details we refer to the books of Gamkrelidze [22], Hermes and LaSalle [24], and Warga [38]. For infinite-dimensional systems, where most of the results concern linear systems, the most general of such results appears to be that of Ahmed [2]. However, the main theorem of that paper (Theorem 5.2) has a serious gap in its proof and, in our opinion, is incorrect. Namely, Ahmed [2, p. 262] claims that $y^n \to y$ strongly in the Lebesgue-Bochner space $L^p(E)$ (we use the notation of [2]). However, such a conclusion is not justified from the properties of the sequence $\{y^n\}_{n\geq 1}$ and additional, considerably stronger conditions are eventually needed to get this

* Received by the editors October 2, 1987; accepted for publication (in revised form) May 24, 1988. This research was supported by National Science Foundation grant DMS-8602313.

† Department of Mathematics, University of California, Davis, California 95616.

strong convergence in $L^p(E)$ (for details, see the results of Brooks and Dinculeanu [8], Castaing [10], and Gutman [23]). After all, it is well known (among people working on differential inclusions; see Aubin and Cellina [4]) that to have a relaxation result (even for finite-dimensional differential inclusions), we must have a Hausdorff-Lipschitz orientor field (unless we are dealing with a semilinear inclusion, in which case the solution admits an integral representation in terms of the evolution operator). Given that a control system, under very general measurability hypotheses, can be written as an equivalent evolution inclusion (deparametrization), to have a relaxation result for the control system, we need the resulting velocity field (i.e., the union of all vector fields over all admissible controls), to be Hausdorff-Lipschitz in the state variable. However, in Ahmed's work [2], which deals with general nonlinear systems, this is not the case. In this paper, by using some recent results on measurable multifunctions obtained in [27], we are able to overcome the difficulties that have appeared in Ahmed's work. In fact, a crucial point in our approach is the relative compactness of the set of original trajectories, which is achieved through appropriate hypotheses on the evolution equation describing the dynamics of the system. Nevertheless, the work of Ahmed [2] is still valuable because it suggests way to address the relaxation problem for infinite-dimensional systems. In this paper, we adopt some of those ideas and techniques which, combined with other tools from different areas, give us two relaxation theorems: one for semilinear and the other for nonlinear infinite-dimensional control systems. We also go beyond Ahmed [2] and study the properties of the set-relaxed trajectories and their dependence on the relaxed controls that generate them. We also prove that, even in the absence of a relaxation result, we can show under mild assumptions on the cost functional that the value of the original optimal control problem equals the value of the one that is relaxed. We conclude with some examples from distributed parameter systems.

2. Preliminaries. Let (Ω, Σ) be a measurable space and X a separable Banach space. Throughout this paper we will be using the following notation:

$$P_{f(c)}(X) = \{A \subseteq X: \text{nonempty, closed, (convex)}\},$$

$$P_{(w)k(c)}(X) = \{A \subseteq X: \text{nonempty, } (w\text{-})\text{compact, (convex)}\}.$$

A multifunction $F: \Omega \to P_f(X)$ is said to be measurable, if for all $z \in X$, $\omega \to d(z, F(\omega)) = \inf\{\|z - x\|: x \in F(\omega)\}$ is measurable. When there is a σ-finite measure $\mu(\cdot)$, with respect to which Σ is complete, the above definition of measurability is equivalent to saying that $\operatorname{Gr} F = \{(\omega, x) \in \Omega \times X: x \in F(\omega)\} \in \Sigma \times B(X)$, where $B(X)$ is the Borel σ-field of X (graph measurability of $F(\cdot)$). We use $S_F^p (1 \leqq p \leqq \infty)$ to denote the set of selectors of $F(\cdot)$ that belong in the Lebesgue-Bochner space $L^p(X)$, i.e., $S_F^p = \{f \in L^p(X): f(\omega) \in F(\omega) \mu\text{-a.e.}\}$. This set may be empty. When $F(\cdot)$ is L^p-integrably bounded (i.e., $F(\cdot)$ is measurable and $\omega \to |F(\omega)\| = \sup\{\|z\|: z \in F(\omega)\} \in L^p\Omega)$), then $S_F^p \neq \varnothing$. Using S_F^1, we can define a set-valued integral for $F(\cdot)$ by setting $\int_\Omega F = \{\int_\Omega f: f \in S_F^1\}$. We also use S_F to denote the set of measurable selectors of $F(\cdot)$.

Let Y, Z be Hausdorff topological spaces and let $G: Y \to 2^Z \setminus \{\varnothing\}$ be a multifunction. We say that $G(\cdot)$ is upper semicontinuous (u.s.c.) (respectively, lower semicontinuous (l.s.c.)), if for all $V \subseteq Z$ open, $G^+(V) = \{y \in Y: G(y) \subseteq V\}$ is open (respectively, $G^-(V) = \{y \in Y: G(y) \cap V \neq \varnothing\}$ is open). Observe that when $G(\cdot)$ is single-valued, then the notions above coincide with the continuity of $G(\cdot)$. If X is a metric space on $P_f(X)$ we can define a metric, known as the Hausdorff metric, by setting: $h(A, B) = \max(\sup[d(a, B), a \in A], \sup[d(b, A), b \in B])$. If X is complete, then so is $(P_f(X), h)$.

Let Z be a separable, complete metric space (i.e., a Polish space) and $B(Z)$ its Borel σ-field. By $M_+^1(Z)$ we will denote the space of probability measures on Z. A

transition probability is a function $\lambda : \Omega \times B(Z) \to [0, 1]$ such that for every $A \in B(Z)$, $\lambda(\cdot, A)$ is Σ-measurable and for every $\omega \in \Omega$, $\lambda(\omega, \cdot) \in M_+^1(Z)$. We use $R(\Omega, Z)$, to denote the set of all transition probabilities from (Ω, Σ, μ) into $(Z, B(Z))$. Following Balder [5] (see also Warga [38]), we can define a topology on $R(\Omega, Z)$ as follows. Let $f : \Omega \times Z \to R$ be a Carathéodory function (i.e., $\omega \to f(\omega, x)$ is measurable, $x \to f(\omega, x)$ is continuous, and $|f(\omega, x)| \leqq \alpha(\omega)$ μ-almost everywhere, with $\alpha(\cdot) \in L^1$) and let $I_f : R(\Omega, Z) \to R$ be defined by $I_f(\lambda) = \int_\Omega \int_Z f(\omega, z)\lambda(\omega)(dz)\, d\mu(\omega)$. The weakest topology on $R(\Omega, Z)$ that makes the above functionals continuous (for any Carathéodory integrand $f(\cdot, \cdot)$) is called the weak topology on $R(\Omega, Z)$. Observe that when Ω is a singleton, then $R(\Omega, Z) = M_+^1(Z)$, and the weak topology just defined is nothing else but the well-known narrow topology on $M_+^1(Z)$ (see Dellacherie and Meyer [17]).

In the rest of this section, for the convenience of the reader, we state without proof (although detailed references are given) some theorems that we will need in the sequel.

We start with the Arzelà-Ascoli Theorem for vector-valued functions. A proof of this result can be found in Carroll [9, Thm. 8.18, p. 34] (see also Lakshmikantham and Leela [28, Thms. 1.15, 1.1.6, p. 5]).

THEOREM 2.1 (Arzelà-Ascoli). *Let Y and Z be Hausdorff topological spaces with Y locally compact. Then $K \subseteq C(Y, Z)$ is relatively compact for the topology of uniform convergence on compacta if and only if K is equicontinuous and for all $y \in Y$, $K(y) = \{f(y) : f \in K\}$ is relatively compact in Z.*

We will use this theorem twice, in the proofs of Theorems 3.1 and 6.1. In the first case $Y = T = [0, b] \subseteq R_+$ and $Z = H$, a separable Hilbert space with the strong (norm) topology. The equicontinuity follows from the fact that the continuous functions in question have an integral representation in terms of an evolution operator $S(t, s)$ that is compact for $t - s > 0$ and hence has nice continuity properties. The pointwise relatively compact range requirement follows from the compactness of $S(t, s)$ for $t - s > 0$ and the Rådström embedding theorem (see below), which tells us that a certain set-valued integral, containing the functions at t translated by $S(t, 0)x_0$, is compact. In the second case (proof of Theorem 6.1), again $Y = T = [0, b]$, while $Z = X_w$ is a separable, reflexive Banach space with weak topology. Here the equicontinuity is a consequence of some a priori estimates that can be deduced from the hypotheses on the data of the problem. On the other hand, the functions are bounded uniformly in t by a constant $\bar{M}$ (see the proof of Theorem 6.1), and in a separable, reflexive Banach space, the closed balls with the weak topology are compact and metrizable. The metrizability then justifies the sequential compactness that we have.

Now let us state the Rådström embedding theorem, which as we say above, will be used in connection with the Arzelà-Ascoli Theorem. The interested reader can find more details in Hiai and Umegaki [25, § 3].

THEOREM 2.2 (Rådström). *Let X be a separable Banach space. The metric space $(P_{kc}(X), h)$ can be embedded as a convex cone in a separable Banach space $\hat{X}$ such that*: (i) *the embedding is isometric*; (ii) *addition in $\hat{X}$ induces addition in $P_{kc}(X)$*; (iii) *multiplication by nonnegative real numbers in $\hat{X}$ induces the corresponding operation in $P_{kc}(X)$.*

This theorem allows us to view integrably bounded multifunctions with values in $P_{kc}(X)$ as $\hat{X}$-valued functions belonging in the Lebesgue-Bochner space $L^1(\hat{X})$. Therefore the set-valued integral of such a multifunction will produce a set in $P_{kc}(X)$.

In some cases we will need to guarantee that the set of admissible controls is nonempty, or to express a selector of the field of velocities as a vector field corresponding

to an admissible control function. This can be done with an application of the so-called "Aumann Selection Theorem." The version presented here is due to Saint-Beuve [33, Thm. 3]. Recall that a Souslin space is a Hausdorff topological space V such that there exists a Polish space W and a continuous map from W onto V (recall that a Polish space is a complete, separable, metrizable space). Clearly every Polish space (for example, a separable Banach space) is Souslin. However, in general, a Souslin space need not be metrizable. Consider, for example, a separable Banach space with the weak topology. This is Souslin but not metrizable.

THEOREM 2.3 (Aumann). *Let (Ω, Σ, μ) be a σ-finite measure space, V a Souslin space, and $F: \Omega \to 2^V \setminus \{\emptyset\}$ a graph-measurable multifunction. Then there exist Σ-measurable functions $f_n: \Omega \to V$ such that $F(\omega) \subseteq \overline{\{f_n(\omega)\}}_{n \geq 1}$ μ-almost everywhere.*

If Σ is μ-complete in the theorem above, then the conclusion will hold for all $\omega \in \Omega$. Finally, if there is no measure μ on (Ω, Σ), then the selectors $f_n(\cdot)$ will be $\hat{\Sigma}$-measurable, where $\hat{\Sigma}$ is the universal σ-field corresponding to Σ.

Another result of measure-theoretic nature that we will need is a "projection theorem," known in the literature as the "Arsenin-Novikov Theorem." The version we present here is due to Dellacherie [16] and Saint-Beuve [34, Thm. 1].

THEOREM 2.4 (Arsenin-Novikov). *Let X, Y be Polish spaces with $B(X)$ and $B(Y)$ the corresponding Borel σ-fields. Let $K \in B(X) \times B(Y)$ be such that for all $x \in X$, $K(x)$ is σ-compact in Y. Then* $\operatorname{proj}_X K \in B(X)$.

Note that instead of $(X, B(X))$, we could have considered any standard measurable space.

In our work (see the proof of Theorem 6.3), $Y = W$, a weakly compact, convex set of a separable Banach space. Recall (see Dunford and Schwartz [19, Thm. 3, p. 434]), that W with the weak topology (denoted by W_w) is metrizable, and hence a Polish space. On the other hand, in that proof we will have $K \in B(X) \times B(W)$. Since $B(W_w) \subseteq B(W)$, to apply the Arsenin-Novikov Theorem, we need to know if equality can hold. This is guaranteed by the following result of Edgar [20, Cor. 2.4]. Recall that a norm $\|\cdot\|$ on a Banach space X is called a "Kadec norm" if and only if the weak and norm topologies of X coincide on $S_X = \{x \in X: \|x\| = 1\}$. We say that X admits a Kadec norm if and only if there is an equivalent norm that is a Kadec norm.

THEOREM 2.5 (Edgar). *Let X be a Banach space that admits a Kadec norm. Let X_w denote the Banach space X with the weak topology. Then $B(X) = B(X_w)$.*

Every weakly compactly-generated Banach space (in particular, every separable Banach space) admits a Kadec norm.

Finally we make a straightforward but nevertheless useful observation. Suppose $T = [0, b] \subseteq R_+$ and Z is a compact Polish space. Then the Carathéodory integrands on $T \times Z$ can be identified with the Lebesgue-Bochner space $L^1(C(Z))$. To see this, associate to each Carathéodory integrand $\phi(\cdot, \cdot)$ the map $t \to \phi(t, \cdot) \in C(Z)$. Now, from the Riesz Representation Theorem we know that $[C(Z)]^* = M(Z)$ is the space of all bounded Borel measures on $B(Z)$. So $M(Z)$ is a separable, dual Banach space and hence has the Radon-Nikodym property. This observation combined with Theorem 1 of Diestel and Uhl [18, p. 98], tells us that $[L^1(C(Z))]^* = L^\infty(M(Z))$. So the weak topology on $R(T, Z)$ coincides with the relative $w^*(L^\infty(M(Z)), L^1(C(Z)))$-topology (see Warga [38]). This fact will be useful in the study of the relaxed control system, where the control functions are transition probabilities.

3. Existence and relaxation results. Let $T = [0, b]$ be closed and bounded. Let H be a separable Hilbert space and X a separable, reflexive Banach space with the following properties: X is dense in H, and the inclusion of X in H is continuous. We

identify H with its dual (pivot space) and denote by X^* the topological dual of X. So $X \hookrightarrow H \hookrightarrow X^*$. By $\|\cdot\|$ (respectively, $|\cdot|$, $\|\cdot\|_*$) we will denote the norm of X (respectively of H, X^*). We also use $(\cdot,\cdot)$ to denote the inner product in H and by $\langle\cdot,\cdot\rangle$ the duality brackets for the dual pair (X, X^*). The two are compatible, i.e., if $x \in X \subseteq H$ and $h \in H \subseteq X^*$, we have $\langle x, h\rangle = (x, h)$. Also as our control space we take Z, a Polish space. We consider the following distributed parameter control system:

$$(*) \qquad \begin{aligned} &\dot{x}(t) + A(t)x(t) = f(t, x(t), u(t)), \\ &x(0) = x_0, \qquad u \in S_U^1. \end{aligned}$$

We will make the following hypotheses concerning the system above:

(H(A)) For $t \in T$, $A(t): X \to X^*$ is such that:
(1) $A(t)(\cdot)$ is linear, monotone;
(2) For every $x \in X$, $t', t \in T$ $\|A(t')x - A(t)x\|_* \leqq k|t' - t|\,\|x\|$, $k > 0$;
(3) $\|A(t)x\|_* \leqq c + c_1\|x\|$ almost everywhere $c_1 > 0$, $c \geqq 0$;
(4) $\langle A(t)x, x\rangle \geqq c_2\|x\|^2$, $c_2 > 0$.

(H(f)) $f: T \times H \times Z \to H$ is a function such that:
(1) $(t, z) \to f(t, x, z)$ is measurable;
(2) $x \to f(t, x, z)$ is continuous;
(3) $|f(t, x, z)| \leqq \alpha(t) + b(t)|x|$ almost everywhere with $\alpha(\cdot)$, $b(\cdot) \in L^2_+$.

(H(U)) $U: T \to 2^Z_{\setminus}\{\varnothing\}$ is graph measurable.

Let $g \in L^2(X^*)$ and consider the following evolution equation:

$$(*)' \qquad \begin{aligned} &\dot{x}(t) + A(t)x(t) = g(t), \\ &\dot{x}(0) = x_0 \in X. \end{aligned}$$

Because of hypothesis (H(A)), from Proposition 5.5.1 of Tanabe [35], we know that $(*)'$ has a unique strong solution belonging in

$$W(T) = \{x(\cdot) \in L^2(X) : \dot{x}(\cdot) \in L^2(X^*)\} \subseteq C(T, H).$$

Furthermore, there exists a strongly continuous evolution operator $S(t, s) \in L(H)$, with respect to which the unique strong solution of $(*)'$ has the following integral representation:

$$x(t) = S(t, 0)x_0 + \int_0^t S(t, s)g(s)\,ds.$$

We make the following hypothesis concerning $S(t, s)$:

(H_c) For all $t > s$, $S(t, s)$ is compact.

Note that because of (H(U)), by Aumann's Selection Theorem (see Theorem 2.3 above), we have that $S_U \neq \varnothing$.

First we will establish the nonemptiness of the set of admissible "state-control" pairs for system $(*)$.

THEOREM 3.1. *If* (H(A)), (H(f)), (H(U)), *and* (H_c) *hold and* $u \in S_U$, *then there exists* $x(u)$, *admissible trajectory of* $(*)$ *corresponding to* $u(\cdot)$.

Proof. First we obtain an a priori bound for the solutions of (∗). Let $x(\cdot)$ be such a strong solution. Since $\|S(t,s)\| \leqq M$ for all $0 \leqq s \leqq t \leqq b$ we have that

$$|x(t)| \leqq M|x_0| + \int_0^t M|f(s, x(s), u(s))|\, ds$$

$$\leqq M|x_0| + \int_0^t M(\alpha(s) + b(s)|x(s)|)\, ds$$

$$\Rightarrow |x(t)| \leqq M(|x_0| + \|\alpha\|_1) + M\int_0^t b(s)|x(s)|\, ds.$$

Applying Gronwall's inequality, for all $t \in T$ we obtain

$$|x(t)| = M(|x_0| + \|\alpha\|_1) \exp M\|b\|_1 = M_2.$$

Now let $\hat{f}: T \times H \times Z \to H$ be defined by

$$\hat{f}(t,x,z) = \begin{cases} f(t,x,z), & |x| \leqq M_2, \\ f(t, (M_2 x/|x|), z), & |x| > M_2. \end{cases}$$

Thus $\hat{f}(t,x,z) = f(t, p_{M_2}(x), z)$, where $p_{M_2}(\cdot)$ is the M_2-radial retraction in H. Recalling that $p_{M_2}(\cdot)$ is Lipschitz continuous, we deduce that $\hat{f}(t,x,z)$ has the same measurability and continuity properties as $f(\cdot,\cdot,\cdot)$, i.e., $(t,z) \to \hat{f}(t,x,z)$ is measurable and $x \to \hat{f}(t,x,z)$ is continuous. Furthermore, $|\hat{f}(t,x,z)| \leqq \varphi(t) = \alpha(t) + b(t)M_2$ almost everywhere, $\varphi(\cdot) \in L^2$.

Let $B(\varphi) = \{g \in L^2(H): |g(t)| \leqq \varphi(t) \text{ a.e.}\} \subseteq L^2(H)$.

Pick $g \in B(\varphi)$ and consider the following evolution equation:

$$(*)(g) \qquad \begin{aligned} &\dot{x}(t) + A(t)x(t) = g(t), \\ &x(0) = x_0. \end{aligned}$$

From Theorem 4.2 of Barbu [6, p. 167], we know that $(*)(g)$ has a unique strong solution in $W(T)$. Let $r: B(\varphi) \to C(T,H)$ be the map that assigns to each $g \in B(\varphi)$ the corresponding unique strong solution of $(*)(g)$. We claim that $r(\cdot)$ is continuous. So let $g_n \xrightarrow{S} g$ in $B(\varphi)$. Then we have

$$r(g_n)(t) = x_n(t) = S(t,0)x_0 + \int_0^t S(t,s)g_n(s)\, ds,$$

$$r(g)(t) = x(t) = S(t,0)x_0 + \int_0^t S(t,s)g(s)\, ds$$

$$\Rightarrow |x_n(t) - x(t)| \leqq M\|g_n - g\|_1 \leqq M\sqrt{b}\|g_n - g\|_2 \to 0$$

$$\Rightarrow x_n \to x \quad \text{in } C(T,H)$$

$$\Rightarrow r(\cdot) \text{ is indeed continuous as claimed.}$$

Now let W be the subset of $C(T,H)$ defined by

$$W = \{r(g) \in C(T,H): g \in B(\varphi)\}.$$

The new claim is that W is compact in $C(T,H)$.

To this end let $x \in W$ and let $t', t \in T$, $t < t'$. We have

$$|x(t') - x(t)| \leqq |S(t',0)x_0 - S(t,0)x_0| + \int_t^{t'} \|S(t',s)\|\,|g(s)|\, ds$$

$$+ \int_0^t \|S(t',s) - S(t,s)\|\,|g(s)|\, ds.$$

Because of the strong continuity of the evolution operator $S(t, s)$, given $\varepsilon > 0$ there exists $\delta_1 > 0$ such that if $|t' - t| < \delta_1$, then

$$|S(t', 0)x_0 - S(t, 0)x_0| < \frac{\varepsilon}{3}.$$

Also

$$\int_t^{t'} \|S(t', s)\| |g(s)| \, ds \leqq M \int_t^{t'} |g(s)| \, ds,$$

and since the Lebesgue integral is absolutely continuous, there exists $\delta_2 > 0$ such that if $|t' - t| < \delta_2$, then

$$M \int_t^{t'} |g(s)| \, ds < \frac{\varepsilon}{3}.$$

Finally, because of hypothesis (H_c), from Proposition 2.1 of [30], we know that $t \to S(t, s)$ is continuous in the operator norm topology, uniformly for all $s \in (0, t)$ such that $t - s$ is bounded away from zero. So let $\delta_3 > 0$ such that

$$\int_{t-\delta_3}^{t'} \|S(t', s) - S(t, s)\| |g(s)| \, ds \leqq 2M \int_{t-\delta_3}^{t} |g(s)| \, ds < \frac{\varepsilon}{6}.$$

Also we can find $\delta_4 > 0$ such that for $|t' - t| < \delta_4$ we have

$$\int_0^{t-\delta_3} \|S(t', s) - S(t, s)\| |g(s)| \, ds < \frac{\varepsilon}{6}.$$

So, finally, for $\delta = \min(\delta_1, \delta_2, \delta_3, \delta_4)$ and for $|t' - t| < \delta$, $x(\cdot) \in W$, we have $|x(t') - x(t)| < \varepsilon \Rightarrow W$ is equicontinuous. Also note that $s \to S(t, s)B(\varphi)(s)$ is a measurable, and, due to hypothesis $(H_{c'})$, a $P_{kc}(H)$-valued multifunction. Thus, using Rädström's Embedding Theorem (see Theorem 2.2 in this paper), we have that for all $x \in W$ and all $t \in T$, $x(t) \in S(t, 0)x_0 + \int_0^t S(t, s)B(\varphi)(s) \, ds \in P_{kc}(H)$. Finally it remains to show that W is closed in $C(T, H)$. Let $x_n \to x$ in $C(T, H)$ $x_n \in W$. Then we have

$$x_n(t) = S(t, 0)x_0 + \int_0^t S(t, s)g_n(s) \, ds, \qquad g_n \in B(\varphi).$$

Note that since $L^2(H)$ is Hilbert, $B(\varphi)$ is sequentially w-compact. So by passing to a subsequence if necessary, we may assume that $g_n \xrightarrow{w} g \in B(\varphi)$ in $L^2(H)$ (Alaoglu's Theorem). Then we have

$$\int_0^t S(t, s)g_n(s) \, ds \xrightarrow{w} \int_0^t S(t, s)g(s) \, ds$$

$$\Rightarrow x(t) = S(t, 0)x_0 + \int_0^t S(t, s)g(s) \, ds$$

$$\Rightarrow x = r(g) \in W$$

$$\Rightarrow W \text{ is closed.}$$

Therefore, invoking the Arzelà-Ascoli Theorem, we get that W is compact in $C(T, H)$. Let $F: W \to L^2(H)$ be defined by

$$F(y)(\cdot) = \hat{f}(\cdot, y(\cdot), u(\cdot)).$$

Exploiting the continuity of $y \to \hat{f}(t, y, z)$, through the dominated convergence theorem we get that $F(\cdot)$ is continuous. Furthermore, note that for every $y \in W$, $F(y) \in B(\varphi)$. Then let $k: W \to W$ be defined by $k = r \circ F$. Clearly $k(\cdot)$ is continuous. Applying Schauder's Fixed-Point Theorem to get $x \in W$ such that $k(x) = x \Rightarrow x(\cdot)$, we solve $(*)(g)$ with $g(t) = \hat{f}(t, x(t), u(t))$. As in the beginning of the proof, through Gronwall's inequality, we get that $|x(t)| \leqq M_2 \Rightarrow \hat{f}(t, x(t), u(t)) = f(t, x(t), u(t)) \Rightarrow x(\cdot)$ is the desired admissible trajectory of $(*)$. □

As we mentioned in the Introduction, to solve optimization problems involving $(*)$ and obtain optimal admissible pairs, we need some kind of convexity hypothesis on the orientor field $f(t, x, U(t))$. Here we drop this convexity hypothesis and instead pass on to a larger system with measure controls, known as "relaxed controls." This then raises the fundamental question of how much we enlarged the set of trajectories of the original system. The next theorem answers this question by stating that this process does not essentially alter the original solution set.

First let us introduce this new, larger system, known as the "relaxed system":

$$(*) \qquad \begin{aligned} &\dot{x}(t) + A(t)x(t) = \int_Z f(t, x(t), z)\lambda(t)(dz), \\ &x(0) = x_0, \qquad \lambda \in S_\Sigma. \end{aligned}$$

Here $\Sigma(t) = \{\lambda \in M_+^1(Z): \lambda(U(t)) = 1\}$ and S_Σ is the set of transition probabilities that are selectors of $\Sigma(\cdot)$. We will denote the set of trajectories of $(*)_r$ by P_r and those of $(*)$ by P. Note that since $\delta(U(t)) \subseteq \Sigma(t)$, we have $P \subseteq P_r$ and if the hypotheses of Theorem 3.1 are satisfied, $P \neq \varnothing \Rightarrow P_r \neq \varnothing$. More specifically, given any relaxed control $\lambda \in S_\Sigma$, if we set $\bar{f}(t, x(t), \lambda(t)) = \int_Z f(t, x(t), z)\lambda(t)(dz)$, then working as in the proof of Theorem 3.1, we can show that there exists a relaxed admissible trajectory $x(\lambda)(\cdot)$, corresponding to λ.

To get a useful relation between P and P_r, we need the following stronger hypotheses.

$(\mathrm{H(f)}_1)$ $f: T \times H \times Z \to H$ is a function such that:

(1) $t \to f(t, x, z)$ is measurable;

(2) $x \to -f(t, x, z)$ is continuous, monotone;

(3) $(x, z) \to f(t, x, z)$ is continuous from $H \times Z$ into H_w;

(4) $|f(t, x, z)| \leqq \alpha(t) + b(t)|x|$ almost everywhere, with $\alpha(\cdot)$, $b(\cdot) \in L_+^2$.

$(\mathrm{H(U)}_1)$ $U: T \to P_f(Z)$ is a measurable multifunction.

THEOREM 3.2. *If hypotheses* (H(A)), $(\mathrm{H(f)}_1)$, $(\mathrm{H(U)}_1)$, *and* $(\mathrm{H_c})$ *hold and Z is a compact Polish space then $\varnothing \neq \bar{P} = P_r$ and the set is convex (the closure is taken in $C(T, H)$).*

Proof. From Theorem 3.1 we know that $P \neq \varnothing$ and clearly $P \subseteq P_r$. Take $y \in P_r$ and for $s \in T$ set

$$K(s) = \{f(s, y(s), u): u \in U(s)\},$$

$$K_r(s) = \left\{\bar{f}(s, y(s), \lambda) = \int_Z f(s, y(s), z)\lambda(dz): \lambda \in \Sigma(s)\right\}.$$

Our claim is that $K_r(s) = \overline{\mathrm{conv}}\, K(s)$. Because Z is compact, it is easy to check that $K_r(s) \in P_{fc}(H)$. Thus, since $\delta(U(s)) \subseteq \Sigma(s)$, we get that $\overline{\mathrm{conv}}\, K(s) \subseteq K_r(s)$. On the other hand, let $v \in K_r(s)$. Then

$$v = \int_Z f(t, y(s), z)\lambda(dz)$$

for some $\lambda \in \Sigma(s)$. From Theorem 12.11 of Choquet [14] (see also Corollary 3 of Balder [5]), we know that there exist $\delta(u_n(s))$ and $\lambda_n \in [0, 1]$ such that $\sum_{k=1} \lambda_k \delta(u_k(s)) = \mu_n \to \lambda$ in the narrow topology. So we have that

$$\int_Z f(s, y(s), z)\mu_n(s)(dz) = \sum_{k=1}^{n} \lambda_k f(s, y(s), u_k(s)) \xrightarrow{w} \int_Z f(s, y(s), z(dz))$$
$$= \bar{f}(s, y(s), \lambda)$$
$$\Rightarrow \overline{\text{conv}}\, K(s) \supseteq K_r(s)$$
$$\Rightarrow \overline{\text{conv}}\, K(s) = K_r(s).$$

Also note that if $\{u_n(\cdot)\}_{n \geq 1}$ is a Castaing representation of the measurable multifunction $U(\cdot)$ (see Wagner [36]), then because of $\text{H(f)}_1(1)$ we have that

$$K(s) = \text{cl}\{f(s, y(s), u_n(s)) : n \geq 1\}.$$

Because of (H(f)_1) (1) and (2), for every $n \geq 1$, $s \to f(s, y(s), u_n(s))$ is measurable $\Rightarrow$ $s \to K(s)$ is a measurable $P_{wk}(H)$-valued multifunction (recall that Z is compact and $f(t, y(t), \cdot)$ is continuous from Z into H_w). Invoking the Krein-Smulian Theorem (see Diestel and Uhl [18, Thm. 11, p. 51]), we deduce that $s \to \overline{\text{conv}}\, K(s)$ is a $P_{wkc}(H)$-valued, measurable multifunction. Now from Corollary II of [27] we know that $\overline{S^1_K}^w = S^1_{K_r}$.

From Proposition 3.1 of [31] we know that $S^1_{K_r}$ is w-compact in $L^1(H)$, and since the space is separable, we deduce that the weak topology on $S^1_{K_r}$ is metrizable. Note that $g(\cdot) = \bar{f}(\cdot, y(\cdot), \lambda(\cdot)) \in S^1_{K_r}$. So according to the above, we can find $g_n \in S^1_K$ such that $g_n \xrightarrow{w} g$, $g_n \in S^1_K$. A simple application of the Aumann Selection Theorem gives us $u_n \in S^1_U$ such that $g_n(s) = f(s, y(s), u_n(s))$ almost everywhere. Let $x_n(\cdot)$ be the original trajectories corresponding to the control functions $u_n(\cdot)$. From the proof of Theorem 3.1, we know that $\{x_n(\cdot)\}_{n \geq 1}$ is relatively compact in $C(T, H)$, and so, by passing to a subsequence if necessary, we may assume that $x_n \to x$ in $C(T, H)$.

Now, note that

$$\frac{d}{dt}|y(s) - x_n(s)|^2 = 2\langle \dot{y}(s) - \dot{x}_n(s), y(s) - x_n(s)\rangle$$
$$= 2\langle -A(s)y(s) + \bar{f}(s, y(s), \lambda(s)) + A(s)x_n(s)$$
$$- f(s, x_n(s), u_n(s)), y(s) - x_n(s)\rangle.$$

Exploiting the monotonicity of $A(s)(\cdot)$, we get

$$\frac{d}{dt}|y(s) - x_n(s)|^2 \leq 2\langle \bar{f}(s, y(s), \lambda(s)) - f(s, x_n(s), u_n(s)), y(s) - x_n(s)\rangle.$$

Integrating both sides we get

$$|y(t) - x_n(t)|^2 \leq 2\int_0^t \langle \bar{f}(s, y(s), \lambda(s)) - f(s, y(s), u_n(s)), y(s) - x_n(s)\rangle\, ds$$
$$+ 2\int_0^t \langle f(s, y(s), u_n(s)) - f(s, x_n(s), u_n(s)), y(s) - x_n(s)\rangle\, ds.$$

Exploiting the dissipativity property of $f(t, \cdot, z)$, we get that the second integral is less or equal to zero. So

$$|y(t) - x_n(t)|^2 \leq 2((g - g_n, y - x_n))$$

where $((\cdot,\cdot))$ denotes the duality brackets for the pair $(L^1(H), L^\infty(H))$. Recall that $g_n \xrightarrow{w} g$ in $L^1(H)$, while $y - x_n \xrightarrow{s} y - x$ in $L^\infty(H)$. Therefore

$$\begin{aligned}
&((g - g_n, y - x_n)) \to 0\\
&\Rightarrow |y(t) - x_n(t)|^2 \to |y(t) - x(t)|^2 = 0 \quad \text{for all } t \in T,\\
&\Rightarrow y = x, \quad \text{i.e.,} \quad y \in \bar{P} \quad \text{(the closure in } C(T, H)).
\end{aligned}$$

Finally we will show that P_r is closed in $C(T, H)$. So let $y_n \to y$ in $C(T, H)$, $y_n \in P_r$. By definition we have

$$y_n(t) = S(t, 0)x_0 + \int_0^t S(t, s)\bar{f}(s, y_n(s), \lambda_n(s))\, ds, \qquad y_n \in S_\Sigma.$$

From Theorem V-1 of Castaing and Valadier [11] we know that S_Σ is w^*-compact in $L^\infty(M(Z))$. Since this topology on S_Σ is metrizable (see Dunford and Schwartz [19, Thm. 1, p. 426]), by passing to a subsequence if necessary, we may assume that $\lambda_n \xrightarrow{w} \lambda$ in $L^\infty(M(Z))$. Then using Theorem 3.1 of Jawhar [26], we get that

$$\begin{aligned}
&y_n(t) \xrightarrow{w} y(t) = S(t, 0)x_0 + \int_0^t S(t, s)\bar{f}(s, y(s), \lambda(s))\, ds, \qquad \lambda \in S_\Sigma\\
&\Rightarrow y \in P_r\\
&\Rightarrow P_r \quad \text{is closed in } C(T, H)\\
&\Rightarrow \bar{P} = P_r \quad \text{in } C(T, H)
\end{aligned}$$

and it is clear that P_r is convex. □

Remarks. Scrutinizing the proof above, we can see that instead of the dissipativity hypothesis on $x \to f(t, x, z)$, we could have assumed that $x \to f(t, x, z)$ is Lipschitz continuous. Also it is clear now that the attainable set of $(*)$ is dense in that of $(*)_r$.

4. Relaxed trajectories. We start with a result describing the dependence of the relaxed trajectories on the relaxed controls that generate them.

Such a continuity result gives us valuable information about the topological structure of the set of relaxed trajectories and is an indispensable tool in establishing the existence of optimal "state-control" pairs in various optimal control problems (see § 5).

From § 3 we know that, given $\lambda \in S_\Sigma$, there exists a unique relaxed trajectory $x(\lambda)$ corresponding to it (uniqueness follows from the dissipativity hypothesis on $f(t, \cdot, u)$; see hypothesis $(\mathrm{H(f)_1})(2)$). In the next theorem we examine the continuity properties of the map $\lambda \to x(\lambda)$ from $S_\Sigma \subseteq R(T, Z)$ into $C(T, H)$.

THEOREM 4.1. *If hypotheses* (H(A)), $(\mathrm{H(f)_1})$, $(\mathrm{H(U)_1})$, *and* $(\mathrm{H_c})$ *hold and Z is compact, then $\lambda \to x(\lambda)$ is continuous from $S_\Sigma \subseteq R(T, Z)$ into $C(T, H)$.*

Proof. By identifying the Carathéodory integrands with the separable Banach space $L^1(C(Z))$, we see that the $R(T, Z)$-topology on S_Σ coincides with the relative $w^*(L^\infty(M(Z)), L^1(C(Z)))$ topology (recall that $(L^1(C(Z))^* = L^\infty(M(Z)))$ and the latter is metrizable). So we will work with sequences. Let $\lambda_n \to \lambda$ in $R(T, Z) \Rightarrow \lambda_n \xrightarrow{w^*} \lambda$. By definition we have

$$x_n(t) = S(t, 0)x_0 + \int_0^t \int_Z S(t, s)f(x, x_n(s), z)\lambda_n(s)(dz)\, ds.$$

Let $h \in H$. We have

$$(h, x_n(t)) = (h, S(t, 0)x_0) + \int_0^t (g(h)(s, x_n(s), \cdot), \lambda_n(s))_0\, ds$$

where $g(h)(s, x_n(s), \cdot) = (h, S(t, s)f(s, x_n(s), \cdot)) \in C(Z)$ and $(\cdot, \cdot)_0$ denotes the duality brackets for the pair $(C(Z), M(Z))$. We may assume that $x_n(\cdot) \to x(\cdot)$ in $C(T, H)$.

Note that $\sup\{|(h, S(t, s)(f(s, x_n(s), z) - f(s, x(s), z)))| : z \in Z\} = |(h, S(t, s)(f(s, x_n(s), z_n) - f(s, x(s), z_n)))|$, for some $z_n \in Z$ depending on (h, t, s). Since Z is compact, by passing to a subsequence if necessary we may assume that $z_n \to z$. Then from hypothesis $(\mathrm{H(f)}_2)(3)$ we have that

$$|(h, S(t, s)(f(s, x_n(s), z_n) - f(s, x(s), z)))| \to 0$$

$$\Rightarrow g(h)(s, x_n(s), \cdot) \to g(h)(s, x(s), \cdot) \quad \text{in } C(Z)$$

$$\Rightarrow g(h)(\cdot, x_n(\cdot), \cdot) \to g(h)(\cdot, x(\cdot), \cdot) \quad \text{in } L^1(C(Z))$$

because of the dominated convergence theorem. Since $\lambda_n \xrightarrow{w^*} \lambda$ in $L^\infty(M(Z)) = [L^1(C(Z))]^*$, we get

$$\int_0^t (g(h)(s, x_n(s), \cdot), \lambda_n(s))_0 \, ds \to \int_0^t (g(h)(s, x(s), \cdot), \lambda(s))_0 \, ds.$$

Thus in the limit we have

$$(h, x(t)) = (h, S(t, 0)x_0) + \left(h, \int_0^t \int_Z S(t, s)f(s, x(s), z)\lambda(s)(dz)\, ds\right)$$

$$\Rightarrow x(t) = S(t, 0)x_0 + \int_0^t S(t, s) \int_Z f(s, x(s), z)\lambda(s)(dz)\, ds$$

$$\Rightarrow x = x(\lambda)$$

$$\Rightarrow \lambda \to x(\lambda) \text{ is continuous on } S_\Sigma \subseteq R(T, Z) \text{ as claimed.} \qquad \square$$

An immediate, interesting consequence of Theorem 4.1 is the following theorem.

THEOREM 4.2. *If the hypotheses of Theorem* 4.1 *hold, then P_r is compact in $C(T, H)$.*

Proof. Recall that S_Σ is w^*-compact in $L^\infty(M(Z))$ and is weakly compact in $R(T, Z)$, and the map $\lambda \to x(\lambda)$ is continuous (Theorem 4.1). Therefore $x(S_\Sigma) = P_r$ is compact in $C(T, H)$.

Remark. We could have deduced Theorem 4.2 from the proofs of Theorems 3.1 and 3.2, where we say that P_r is closed in $C(T, H)$ and lives inside a compact subset of $C(T, H)$.

5. Optimal control problems. As we have already mentioned, the introduction of the larger relaxed system, guarantees the existence of an optimal solution. This is illustrated by the following general result.

Consider the relaxed control system $(*)_r$, with the following cost functional:

$$J_r(x, \lambda) = \int_0^b \int_Z L(t, x(t), z)\lambda(t)(dz)\, dt.$$

We will make the following hypotheses concerning the integrand $L(\cdot, \cdot, \cdot)$:

(H(L)) $L: T \times X \times Z \to \bar{R} = R \cup \{+\infty\}$:

(1) $(t, x, z) \to L(t, x, z)$ is measurable,

(2) $(x, z) \to L(t, x, z)$ is l.s.c.,

(3) $\psi(t) \leqq L(t, x, z)$ almost everywhere, with $\psi(\cdot) \in L^1$.

Let $m_r = \inf\{J_r(x, u) : (x, u) \in A_r(x_0)\}$, where $A_r(x_0)$ is the set of relaxed admissible pairs.

THEOREM 5.1. *If hypotheses* (H(A)), (H(f)$_1$), (H(U)$_1$), (H$_c$), *and* (H(L)) *hold and* Z *is compact, then there exists* $(x, \lambda) \in A_r(x_0)$ *such that* $m_r = J_r(x, \lambda)$.

Proof. Let $\{(x_n, \lambda_n)\}_{n \geq 1}$ be a minimizing sequence in $A_r(x_0)$. Recall that S_Σ is w^*-compact in $L^\infty(M(Z))$ and because of the metrizability of this relative w^*-topology ($(L^1(C(Z))$ being separable), by passing to a subsequence if necessary, we may assume that $\lambda_n \overset{w^*}{\to} \lambda$. Identifying as before $L^1(C(Z))$ with the space of Carathéodory integrands, we see that $\lambda_n \to \lambda$ in $R(T, Z)$ with the weak topology. Applying Theorem 4.1 we get that $x(\lambda_n) = x_n \to x(\lambda) = x$ in $C(T, H)$. Recalling that every lower semicontinuous measurable integrand is the limit of an increasing sequence of Carathéodory integrands, from the definition of the weak topology on $R(T, Z)$ (see also Balder [5] and Jawhar [26]), we have that

$$\underline{\lim}\, J_r(x_n, \lambda_n) \geq \int_0^b \int_Z L(t, x(t), z)\lambda(t)(dz)\, dt$$

$$\Rightarrow m_r \geq J(x, \lambda).$$

But $(x, \lambda) \in A_r(x_0)$. Therefore, $m_r = J(x, \lambda)$ as claimed by the theorem. □

Remark. If $J(x, u) = \int_0^b L(t, x(t), u(t))\, dt$ is the cost functional for the original system and $m = \inf\{J(x, u) : (x, u) \in A(x_0)\}$, the value of the corresponding optimization problem, then if the cost functional is upper semicontinuous on $C(T, H) \times R(T, Z)$, the space $R(T, Z)$ endowed the weak topology, then, because of the density result proved in Theorem 3.2, we have that $m = m_r$.

Here is a more general situation, for which this equality of the values of the two problems is still true.

We will need the following stronger hypotheses on $L(\cdot, \cdot, \cdot)$.

(H(L)$_1$) $L : T \times X \times Z \to R$ is an integrand such that:
(1) $t \to L(t, x, z)$ is measurable;
(2) $(x, z) \to L(t, x, z)$ is continuous;
(3) $|L(t, x, z)| \leq \psi_1(t) + \psi_2(t) r(x)$ almost everywhere, with $\psi_1(\cdot), \psi_2(\cdot) \in L^1$ and $r : X \to R$ bounded.

THEOREM 5.2. *If hypotheses* (H(A)), (H(f)$_1$), (H(U)$_1$), (H$_c$), *and* (H(L)$_1$) *hold and* Z *is compact, then there exists* $(x, \lambda) \in A_r(x_0)$ *such that* $J_r(x, \lambda) = m_r$ *and* $m = m_r$.

Proof. From Theorem 5.1 we know that there exists $(x, \lambda) \in A_r(x_0)$ such that

$$m_r = J_r(x, \lambda).$$

Note that we always have $m_r \leq m$. On the other hand, using Corollary 3 of Balder [5], we can find $u_n \in S_U$ such that $\delta(u_n) \to \lambda$ in $R(T, Z)$ with the weak topology. Then from hypothesis (H(L)$_1$) we have that

$$\int_0^b L(t, x(t), u_n(t))\, dt \to \int_0^b \int_Z L(t, x(t), z)\lambda(t)(dz)\, dt = m_r$$

$$\Rightarrow m \leq m_r$$

$$\Rightarrow m = m_r.$$ □

We conclude this section with a time optimal control problem. So let $V : T \to P_f(H)$ be a moving target set that is u.s.c. from T into H_w. Our goal is to reach $V(\cdot)$ in minimum time. We will make the following controllability type hypothesis: $\{t \in T : V(t) \cap P_r(t) \neq \emptyset\} \neq \emptyset$.

THEOREM 5.3. *If the hypotheses of Theorem* 4.2 *hold, then there exists a time-optimal relaxed control.*

Proof. Let $\tau = \inf\{t \in T : V(t) \cap P_r(t) \neq \varnothing\}$. Let $t_n \downarrow \tau$. Then there exist $x_n \in P_r$ such that $x_n(t_n) \in V(t_n) \cap P_r(t_n)$. Using Theorem 4.2 and passing to a subsequence, we may assume that $x_n \to x \in P_r \Rightarrow x_n(t_n) \to x(\tau) \in P_r(\tau)$. Also $x(\tau) \in w - \overline{\lim}\, V(t_n) \subseteq V(\tau) \Rightarrow x(\tau) \in V(\tau) \cap P_r(\tau)$. Let $\lambda \in S_\Sigma$ be the relaxed control that generates x. This is the desired relaxed time-optimal control. □

6. Nonlinear evolution equations. We can also have a relaxation result for evolution equations that is more general than the results for semilinear equations considered in the previous sections.

Let (X, H, X^*) be a Gelfand triple of spaces as before and let Z be a separable Banach space (control space). Consider the following nonlinear, infinite-dimensional control system:

$$(**) \qquad \begin{aligned} &\dot{x}(t) + A(t, x(t), u(t)) = 0, \\ &x(0) = x_0, \qquad u \in S^1_U. \end{aligned}$$

By P we will denote the set of admissible trajectories of (**).

We will need the following hypotheses.

$(\mathrm{H(A)}_1)$ $A : T \times X \times Z \to X^*$ is an operator such that:

(1) $t \to A(t, x, z)$ is measurable;

(2) $(x, z) \to A(t, x, z)$ is sequentially weakly continuous;

(3) $x \to A(t, x, z)$ is monotone;

(4) $\|A(t, x, z)\|_* \leqq \alpha(t) + c_1\|x\|^{p-1}$ almost everywhere, with $\alpha \in L^q$,

$$c_1 > 0, \quad 1 < p < \infty, \quad 1/p + 1/q = 1;$$

(5) $\langle A(t)x, x\rangle \geqq c_2\|x\|^p$ almost everywhere, with $c_2 > 0$.

$(\mathrm{H(U)}_2)$ $U : T \to P_{wkc}(X)$ is integrably bounded.

The first theorem describes the topological properties of P. Its proof is based on a pair of simple lemmata that produce some a priori bounds implied by hypotheses $(\mathrm{H(A)}_1)$.

LEMMA 6.1. *If hypotheses* $(\mathrm{H(A)}_1)$ *and* $(\mathrm{H(U)}_2)$ *hold and* $x_0 \in X$, *then* P *is relatively weakly compact in* $L^p(X)$.

Proof. From Barbu [6], we know that $P \neq \varnothing$. Let $x(\cdot) \in P$. We have

$$\langle \dot{x}(t), x(t)\rangle = \langle -A(t, x(t), u(t)), x(t)\rangle \quad \text{a.e.}, \qquad u \in S^1_U$$

$$\Rightarrow \frac{d}{dt}|x(t)|^2 + 2\langle A(t, x(t), u(t)), x(t)\rangle = 0 \quad \text{a.e.}$$

Because of hypotheses $(\mathrm{H(A)}_1)(1)$ and (2), $t \to \langle A(t, x(t), u(t)), x(t)\rangle$ is measurable. So integrating and using hypothesis $(\mathrm{H(A)}_1)(5)$ we get

$$|x(b)|^2 - |x_0|^2 + c_2\|x\|_p^p \leqq 0$$

$$\Rightarrow P \text{ is bounded in } L^p(X).$$

But $L^p(X)$ is reflexive, since X is. So from Alaoglu's theorem we get the conclusion of the lemma. □

As before, we can have a compactness result for the set of admissible velocities of (**). Let us denote this set by the suggestive notation $\dot{P}$.

LEMMA 6.2. *If hypotheses* $(H(A)_1)$ *and* $(H(U)_2)$ *hold and* $x_0 \in X$, *then* $\dot{P}$ *is relatively weakly compact in* $L^q(X^*)$.

Proof. Recall that $P \subseteq W(T) \Rightarrow \dot{P} \subseteq L^q(X^*)$. Let $v \in L^p(X)$. Then for $x \in P$ we have

$$\int_0^b \langle \dot{x}(t), v(t) \rangle \, dt = \int_0^b -\langle A(t, x,(t), u(t)), v(t) \rangle \, dt$$

$$\leqq \int_0^b \|A(t, x(t), u(t))\|_* \|v(t)\| \, dt.$$

Applying Hölder's inequality and using hypothesis (H(A))(4), we get

$$\int_0^b \|A(t, x(t), u(t))\|_* \|v(t)\| \, dt \leqq \left[\int_0^b \|A(t, x(t), u(t))\|_*^q \, dt \right]^{1/q} \cdot \|v\|_p$$

$$\leqq \left[\int_0^b (\alpha(t) + c_1 \|x(t)\|^{p-1})^{p/(p-1)} \right]^{1/q} \cdot \|v\|_p$$

$$\leqq \left(\|\alpha\|_q^q + c_1 \|x\|_p^p \right)^{1/q} \cdot \|v\|_p.$$

But from Lemma 3.1 we know that $\sup \{\|x\|_p : x \in P\} \leqq M < \infty$. Hence we get

$$((\dot{x}, v)) \leqq [\|\alpha\|_q^q + c_1 M^p]^{1/q} \|v\|_p$$

where $((\cdot, \cdot))$ denotes the duality brackets between $L^p(X)$ and $L^q(X^*)$. Therefore, we finally have that

$$\sup \{((\dot{x}, v)) : \|v\|_p \leqq 1\} = \|\dot{x}\|_q \leqq [\|\alpha\|^q + c_1 M^p]^{1/q} = M_0 < \infty$$

$$\Rightarrow \dot{P} \text{ is bounded in } L^q(X^*)$$

$$\Rightarrow \dot{P} \text{ is relatively weakly compact in } L^q(X^*). \qquad \square$$

Now we are ready for the theorem on the topological properties of P. A result of Ahmed and Teo [3] is similar but concerns a smaller class of nonlinear systems with more restrictive hypotheses.

THEOREM 6.1. *If hypotheses* $(H(A)_1)$ *and* $(H(U)_2)$ *hold and* $x_0 \in X$ *then* P *is relatively sequentially compact in* $C(T, X_w)$.

Proof. Let $R = \{y \in L^q(X^*) : \int_A y(s) \, ds \in X \text{ for all } A \subseteq T, \text{ Lebesgue measurable}\}$. Clearly R is a linear subspace of $L^q(X^*)$. Let $\{y_n\}_{n \geqq 1} \subseteq R$, $y_n \xrightarrow{s} y$ in $L^q(X^*)$.

Then for every $x \in X$ we have

$$\int_0^b \langle \chi_A(s)x, y_n(s) \rangle \, ds \to \int_0^b \langle \chi_A(s)x, y(s) \rangle \, ds,$$

$$\left\langle x, \int_A y_n(s) \, ds \right\rangle \to \left\langle x, \int_A y(s) \, ds \right\rangle.$$

Since $X \hookrightarrow H \hookrightarrow X^*$ with all injections continuous and dense, we deduce that $\int_A y_n(s) \, ds \xrightarrow{w} \int_A y(s) \, ds$ in X. But X, being reflexive, is w-complete. So for all $A \subseteq T$ Lebesgue measurable, we have $\int_A y(s) \, ds \in X \Rightarrow y \in R \Rightarrow R$ is a reflexive, separable Banach space in $L^q(X^*)$.

Next, for $A \subseteq T$ Lebesgue measurable, consider the linear operator $K(A) : R \to X$ defined by

$$K(A)(y) = \int_0^b \chi_A(t) y(t) \, dt.$$

For every $x^* \in X^*$, we see that $y \to \langle x^*, K(A)(y)\rangle$ is a continuous linear functional on R. Thus we can find $g(x^*)(\cdot) \in L^p(X)$ such that

$$\langle x^*, K(A)(y)\rangle \int_A \langle g(x^*)(s), y(s)\rangle \, ds \leqq \left[\int_A \|g(x^*)(s)\|^p \, ds\right]^{1/p} \|y\|_p.$$

Also, since $x \in P$, is X^*-absolutely continuous and so from Diestel and Uhl [18, p. 217], we have that

$$x(t+h) - x(t) = \int_t^{t+h} \dot{x}(s) \, ds$$

and because $x_0 \in X$, we get $x \in R$. Then if we set $A = [t, t+h]$ we have

$$\begin{aligned}\langle x^*, x(t+h) - x(t)\rangle &= \langle x^*, K(A)(\dot{x})\rangle \\ &\leqq \left[\int_t^{t+h} \|g(x^*)(s)\|^p \, ds\right]^{1/p} \|\dot{x}\|_q\end{aligned}$$

and since, by Lemma 6.2, $\dot{P}$ is $L^q(X^*)$-bounded, we deduce that P is w-equicontinuous.

Furthermore, for every $t \in T$ and every $x \in P$, we have

$$|\langle x^*, x(t)\rangle| \leqq \|x_0\| \|x^*\| + \left[\int_0^b \|g(x^*)(s)\|^p \, ds\right]^{1/p} |\dot{P}| \leqq \bar{M}.$$

From the uniform boundedness principle, we get that for all $t \in T$ and all $x \in P$, $x(t) \in B(0, \bar{M}) = \{z \in X : \|z\| \leqq M\}$, which is w-compact. So invoking the Arzelà-Ascoli Theorem (see Theorem 2.1 in this paper and Theorem 1.1.6 of [28]), we have that P is relatively sequentially compact in $C(T, X_w)$. □

To system (**) we associate the following relaxed system:

$$(**) \qquad \begin{aligned} &\dot{x}(t) + \int_Z A(t, x(t), z)\lambda(t)(dz) = 0, \\ &x(0) = x_0, \qquad \lambda \in S_\Sigma \end{aligned}$$

where as before, $\Sigma(t) = \{\lambda \in M_+^1(Z) : \lambda(U(t)) = 1\}$ and S_Σ are the transition probabilities that are selectors of $\Sigma(\cdot)$.

We will need the following stronger version of hypothesis $(\mathrm{H(U)}_2)$.

$(\mathrm{H(U)}_3)$ $U : T \to P_{fc}(Z)$ is measurable and $U(t) \subseteq W$ almost everywhere with $W \in P_{wkc}(Z)$.

Recall (see Dunford and Schwartz [9, Thm. 3, p. 434]) that the weak topology on W is metrizable. So W with the weak topology is a compact Polish space.

By P_r we will denote the set of trajectories of $(**)_r$. Clearly $P \subseteq P_r$. As we did for P, we can have that P_r is relatively sequentially compact in $C(T, X_w)$. In fact we can say more.

THEOREM 6.2. *If hypotheses* $(\mathrm{H(A)}_1)$ *and* $(\mathrm{H(U)}_3)$ *hold and* $x_0 \in X$, *then* P_r *is sequentially compact in* $C(T, X_w)$.

Proof. It suffices to show that P_r is sequentially closed in $C(T, X_w)$. Let $x_n \to x$ in $C(T, X_w)$ with $x_n \in P_r$. We have

$$\dot{x}_n(t) + \int_Z A(t, x_n(t), z)\lambda_n(t)(dz) = 0, \qquad \lambda_n \in S_\Sigma.$$

Since W with the weak topology is a compact Polish space, Theorem V-2 of Castaing and Valadier [11] tells us that S_Σ is w^*-compact in $L^\infty(M(Z))$ (recall that $L^\infty(M(Z)) = (L^1(C(Z)))^*$). So by passing to a subsequence we may assume that $\lambda_n \xrightarrow{w} \lambda$. Then we have

$$x_n(t) - x_0 + \int_0^t \int_W A(s, x_n(s), z)\lambda_n(s)(dz)\, ds$$

$$\xrightarrow{w} x(t) - x_0 + \int_0^t \int_W A(s, x(s), z)\lambda(s)(dz)\, ds$$

$$\Rightarrow x \in P_r$$

$$\Rightarrow P_r \text{ is sequentially compact in } C(T, X_w). \qquad \square$$

Next we prove a relaxation result, involving the sets P and P_r. For this we will need the following stronger version of hypothesis $(\mathrm{H(A)}_1)$.

$(\mathrm{H(A)}_2)$ $A(t, x, z) = A(t, x) - f(t, x, z)$ where $A: T \times X \to X^*$ satisfies $(\mathrm{H(A)}_1)$ (without z) and $f: T \times H \times Z \to H$ satisfies $(\mathrm{H(f)}_1)(1)$, (3), (4) and

(6) $|f(t, x', z) - f(t, x, z)| \leqq k(t)|x' - x|$ almost everywhere, with $k(\cdot) \in L^1_+$.

Also recall that if $f \in L^1(H)$, the weak norm of $f(\cdot)$ is defined by $\|f\|_w = \sup\{\|\int_t^{t'} f(s)\, ds\| : t', t \in T\}$.

THEOREM 6.3. *If hypotheses* $(\mathrm{H(A)}_2)$ *and* $(\mathrm{H(U)}_3)$ *hold,* $x_0 \in X$, *the embedding* $X \hookrightarrow H$ *is compact and* Z *is as before a separable Banach space, then* $\varnothing \neq \bar{P} = P_r$ *and the set is convex* (*the closure is taken in* $C(T, X_w)$).

Proof. Let $x \in P_r(x_0)$ and let $\varepsilon > 0$. Set $F(t, x) = f(t, x, U(t))$ and $F_r(t, x) = \int_w f(t, x, z)\Sigma(t)(dz)$.

Since $\overline{S_{\delta(U(\cdot))}}^{w^*} = S_\Sigma$ in $L^\infty(M(Z))$, working as in the proof of Theorem 3.2, we can show that $F_r(t, x) = \overline{\mathrm{conv}}\, F(t, x)$. Also by definition we have

$$\begin{aligned} \mathrm{Gr}\, F &= \{(t, x, y) \in T \times H \times H : y \in F(t, x)\} \\ &= \{(t, x, y) \in T \times H \times H : y = f(t, x, u),\ u \in U(t)\}. \end{aligned}$$

Set

$$k(t, x, y, u) = y - f(t, x, u) \text{ and } l(t, x, y, u) = d(u, U(t)).$$

Then

$$\begin{aligned} \mathrm{Gr}\, F &= \{(t, x, y) \in T \times H \times H : k(t, x, y, u) = 0,\ l(t, x, y, u) = 0\} \\ &= \mathrm{proj}_{T \times H \times H}\, [(t, x, y, u) : k(t, x, y, u) = 0,\ l(t, x, y, u) = 0]. \end{aligned}$$

Note that both k and l are $B(T) \times B(H) \times B(H) \times B(W)$ measurable. From Theorem 2.5 of this paper (Edgar's Theorem), we know that $B(Z) = B(Z_w)$, where Z_w is the Banach space Z with the weak topology. Hence $B(Z) \cap W = B(Z_w) \cap W$, which implies that $B(W) = B(W_w)$ (again W_w is the set W with the relative weak topology). But recall that W_w is a compact Polish space. So applying Theorem 2.4 (the Arsenin-Novikov Theorem), we get

$$\mathrm{proj}_{T \times H \times H}\, [(t, x, y, u) : k(t, x, y, u) = 0,\ l(t, x, y, u) = 0] \in B(T) \times B(H) \times B(H)$$

$$\Rightarrow \mathrm{Gr}\, F \in B(T) \times B(H) \times B(H).$$

So $F(\cdot,\cdot)$ is graph measurable. This allows us to apply Theorem 2 of Chuong [15] and get that $S^1_{F(\cdot,x(\cdot))}$ is dense in $S^1_{\text{conv } F(\cdot,x(\cdot))} = S^1_{F_r(\cdot,x(\cdot))}$ for the weak norm $\|\cdot\|_w$ introduced earlier. Hence, for every $n \geqq 1$, we can find $f_n \in S^1_{F(\cdot,x(\cdot))}$ such that $\|\dot{g} - f_n\|_w \to 0$, where $g(t) = \dot{x}(t) + A(t, x(t))$. Now consider the multifunction $L_n : T \to 2^Z$ $n \geqq 1$ defined by

$$L_n(t) = \{u \in U(t) : f_n(t) = f(t, x(t), u)\}.$$

From the definition of $F(\cdot,\cdot)$ it is clear that for all $t \in T$, $L_n(t) \neq \varnothing$. Let $\{x_m\}_{m \geqq 1}$ be dense in X and consider the following functions:

$$h^n_m(t, u) = (x_m, f_n(t) - f(t, x(t), u)).$$

Because of hypotheses $(\mathrm{H(f)}_1)$ (1) and (3) we see that for every $m \geqq 1$, $t \to h^n_m(t, u)$ is measurable and $u \to h^n_m(t, u)$ is continuous. Hence we deduce that $(t, u) \to h^n_m(t, u)$ is jointly measurable and so we have that

$$\{(t, u) \in T \times Z : h^n_m(t, u) = 0\} \in B(T) \times B(Z) \quad \text{for every } m \geqq 1.$$

Now observe that

$$\text{Gr } L_n = \left[\bigcap_{m \geqq 1} \{(t, u) \in T \times Z : h^n_m(t, u) = 0\} \right] \cap \text{Gr } U.$$

Recalling that $\text{Gr } U \in B(T) \times B(Z)$ (hypothesis $(\mathrm{H(U)}_3)$), we conclude that for every $n \geqq 1$, we have

$$\text{Gr } L_n \in B(T) \times B(Z).$$

So we can apply Theorem 2.3 (Aumann's Selection Theorem) to find $u_n : T \to Z$ measurable, $n \geqq 1$, such that $u_n(t) \in L_n(t)$ almost everywhere. So we have that $f_n(t) = f(t, x(t), u_n(t))$ almost everywhere with $u_n \in S^1_U$.

Let $y_n(\cdot)$ be the unique strong solution of the original control system $(**)$ corresponding to the admissible control $u_n(\cdot)$. From Theorem 6.1 we know that P is relatively sequentially compact in $C(T, X_w)$. So by passing to a subsequence, if necessary, we may assume that $y_n \to y \in \bar{P}$ in $C(T, X_w)$. Then for any $x \in X$ we have

$$\begin{aligned}
\frac{d}{dt}|x(t) - y_n(t)|^2 &= 2\langle \dot{x}(t) - \dot{y}_n(t), x(t) - y_n(t) \rangle \\
&\leqq 2\langle -A(t, x(t)) + g(t) + A(t, y_n(t)) - f(t, y_n(t), u_n(t), x(t) - y_n(t) \rangle \\
&\Rightarrow |x(t) - y_n(t)|^2 \\
&\leqq \int_0^t (g(s) - g_n(s), x(s) - y_n(s))\, ds \int_0^t k(s)|x(s) - y_n(s)|^2\, ds.
\end{aligned}$$

We know that $y_n \to y$ in $C(T, X_w)$. So $y_n(t) \xrightarrow{w} y(t)$ in X, and since by hypothesis X imbeds compactly into H, we get that $y_n(t) \xrightarrow{s} y(t)$ in H. Also, by construction, $\|g - f_n\|_w \to 0$. Therefore $f_n \xrightarrow{w} g$ in $L^2(H)$ and in the limit as $n \to \infty$ we obtain

$$|x(t) - y(t)|^2 \leqq \int_0^t k(s)|x(s) - y(s)|^2\, ds.$$

Invoking Gronwall's inequality, we get that $x = y$. So we have that x belongs in the closure of P in $C(T, X_w)$. Since $x \in P_r$ is arbitrary we have that $P_r \subseteq \bar{P}$. On the other hand, $P \subseteq P_r$, and by Theorem 6.2, P_r is sequentially compact in $C(T, X_w)$. Therefore we conclude that $\bar{P} = P_r$ as claimed in our theorem. □

Remark. From the proof above it is clear that the Lipschitz condition in the state variable of the vector field $A(t, x, u)$ is essential in obtaining the density result. There is also the counterexample due to Plis (see Aubin and Cellina [4]), showing that even for differential inclusions in R^2, to have a relaxation result we need a Lipschitz hypothesis.

The relaxation result tells us that "essentially" we can have the same attainable set by economizing on the set of controls.

7. Examples. In this section we present two examples from control systems governed by partial differential equations that illustrate the applicability of our results.

Example 1. Let W be an open domain in R^n with smooth boundary $W = G$ and let $T = [0, b]$. On $T \times W$ we consider the following distributed parameter control system:

$$(***)_1 \qquad \frac{\partial x(t,y)}{\partial y} - \sum_{k=1}^{n} \frac{\partial}{\partial y_k}\left(p(t,y) \frac{\partial x(t,y)}{\partial y_k} \right) = f(t, x(t,y), u(t,y)),$$

$x(t, y) = 0$ on $T \times \Gamma$, $x(0, y) = x_0(y)$ on $\{0\} \times W$, $|u(t, y)| \leqq c(t)$ almost everywhere.

Here $p: T \times \bar{W} \to R_+$ is k-Lipschitz in the t-variable, C^1- in the y-variable and $t \to \|p(t, \cdot)\|_\infty \in L^\infty_+$. Also $c: T \to R_+$ is measurable.

Let $X = H^1_0(W)$, $H = L^2(W)$, $X^* = H^{-1}(W) = (H^1_0(W))^*$.

For $t \in T$, let $A(t): X \to X^*$ be the linear operator defined through the Dirichlet form

$$\alpha(t, x, v) = \sum_{k=1}^{n} \int_W p(t, y) \frac{\partial x}{\partial y_k} \frac{\partial v}{\partial y_k} dy$$

y setting $\langle A(t)x, v \rangle = \alpha(t, x, v)$ for all $x, v \in X$.

Because of our hypothesis on $p(\cdot, \cdot)$, we have that $t \to A(t)x$ is k-Lipschitz for $x \in X$. Also for $x, v \in X$, we have

$$\begin{aligned} \langle A(t)x - A(t)v, x - v \rangle &= \alpha(t, x, x - v) - \alpha(t, v, x - v) \\ &= \int_W \sum_{k=1}^{n} p(t, y)\left(\frac{\partial x}{\partial y_k} - \frac{\partial v}{\partial y_k} \right)\left(\frac{\partial x}{\partial y_k} - \frac{\partial v}{\partial y_k} \right) dy \geqq 0. \end{aligned}$$

Furthermore, for $x \in X$ we have

$$\begin{aligned} \|A(t)x\|_* &= \sup \{\langle A(t)x, v \rangle : \|v\| \leqq 1\} \\ &= \sup \{\alpha(t, x, v) : \|v\| \leqq 1\} \\ &= \sup \left\{ \int_W \sum_{k=1}^{n} p(t, y) \frac{\partial x}{\partial y_k} \frac{\partial v}{\partial y_k} dy : \|v\| \leqq 1 \right\}. \end{aligned}$$

Invoking the Cauchy and Poincaré inequalities, we finally have that

$$\|A(t)x\|_* \leqq \|p(t)\|_\infty \|x\|.$$

Finally if $x \in X$, through Poincare's inequality we deduce that

$$\langle A(t)x, x\rangle = \int_W \sum_{k=1}^{n} p(t, y)\left|\frac{\partial x}{\partial y_k}\right|^2 dy \geqq \lambda \|x\|^2.$$

From all of the above, we see that hypothesis $(\mathrm{H(A)}_1)$ is satisfied.

Next assume that $f: T \times \bar{W} \times R \times R^m \to R$ is a function such that it is measurable in $(t, y, u) \in T \times \bar{W} \times R^m$ and continuous in $x \in R$ (so jointly measurable in all variables). Also assume that

$$|f(t, y, x, u)| \leqq \alpha(t, y) + b(y)|x| \quad \text{a.e.}$$

where $\alpha(\cdot, \cdot) \in L^2(T \times W)$ and $b \in L^\infty(W)$. Let $F: T \times L^2(W) \times L^2_m(W) \to L^2(W)$, be defined by

$$F(t, x, u)(y) = f(t, y, x(y), u(y)), \qquad y \in W.$$

From Krasnoselski's Theorem we know that $x \to F(t, x, u)$ is continuous, while $(t, u) \to F(t, x, u)$ is measurable. Furthermore we have

$$|F(t, x, u)| \leqq \alpha(t, \cdot)\|_{L^2(W)} + \|b\|_{L^2(W)}\|x\|_{L^2(W)}.$$

So $F(\cdot, \cdot, \cdot)$ defined as above satisfies hypothesis (H(f)).

Next let $U: T \to P_f(H)$ be defined by

$$U(t) = \{u \in L^2(W): |u(z)| \leqq c(t) \text{ a.e.}\}.$$

Clearly, since $c(\cdot)$ is measurable, $U(\cdot)$ is also measurable. Thus hypothesis (H(U)) is satisfied.

Finally, note that the family of linear operators $\{A(t): t \in T\}$ generates an evolution operator $S(t, s)$ that is compact for $t - s > 0$ (see Martin [29] or Pavel [32]). Therefore hypothesis $(\mathrm{H_c})$ is satisfied. Then Theorem 3.1 gives us the existence of admissible pairs.

If $f(t, y, x, u) = f_0(t, y, x) \cdot u$, where $f_0: T \times \bar{W} \times R \to R^m$ is measurable in (t, y), continuous, dissipative in x, and $|f_0(t, y, x)u| \leqq \alpha(t, y) + b(y)|x|$ almost everywhere as above, then the Nemitsky operator $F(t, x, u)(\cdot) = f_0(t, \cdot, x(\cdot)) \cdot u(\cdot)$ satisfies hypothesis $(\mathrm{H(f)}_1)$. Also if $c(t) = c$ for all $t \in T$, then for all $t \in T$ $U(t) \subseteq \overline{B(0, c\lambda(W))}$ is equal to a ball of radius $c\lambda(W)$ in $L^2(W)$, which is compact-metrizable in the weak topology. Hence if $Z = \overline{B(0, M)}$, then all hypotheses of Theorem 3.2 are satisfied and we have that the set of trajectories of the original system (***) are dense in $C(T, H)$ in those of the convexified (relaxed) system (the one with orientor field $F_r(t, x) = \overline{\text{conv}}\, F(t, x)$). Also, Theorems 4.1 and 4.2 are satisfied and we have information about the properties of the set of relaxed trajectories. Furthermore, the results on the optimal control problems are also valid.

Example 2. In this example, we consider the following nonlinear distributed parameter system (we use the multi-index notation):

$$\begin{aligned}
&\frac{\partial x(t, y)}{\partial y} + \sum_{|\alpha| \leqq m-1} (-1)^{|\alpha|} D^\alpha A_\alpha(t, y, x, \cdot\, Dx, \cdots, Dx^{m-1}) = B(t)u(t, y),\\
&\qquad D^\beta x = 0 \quad \text{on } T \times \Gamma, \qquad |\beta| \leqq m-1,\\
(***)_2 \qquad &\qquad x(0, y) = x_0(y),\\
&\qquad u(t, \cdot) \in V \quad \text{a.e.} \quad \text{with } V \subseteq L^2(W), \text{ } w\text{-compact.}
\end{aligned}$$

As in the previous example, W is a bounded domain in R^n with regular boundary $\partial W = \Gamma$.

For the functions A_α we assume the following:

(1) $(t, y) \to A_\alpha(t, y, r_0, \cdots, r_{m-1})$ is measurable;

(2) $(r_0, \cdots, r_{m-1}) \to A_\alpha(t, y, r_0, \cdots, r_{m-1})$ is continuous;

(3) $|A_\alpha(t, y, r)| \leqq g(t, y) + c \sum_{k=0}^{m-1} |r_k|$ almost everywhere $g \in L^2(T \times W)$; $r = (r_1, \cdots, r_{m-1})$.

(4) $\sum_{|\alpha| \leqq m-1} (A_\alpha(t, y, r) - A_\alpha(t, y, v), r_\alpha - v_\alpha) \geqq 0$.

To the differential operator in divergence form

$$A(t)x = \sum_{|\alpha| \leqq m} (-1)^{|\alpha|} D^\alpha A_\alpha(t, y, x, Dx, \cdots, D^{m-1}x)$$

we associate the following Dirichlet form:

$$p(t, x, v) = \sum_{|\alpha| \leqq m-1} \int_W (A_\alpha(t, y, x, Dx, \cdots, D^{m-1}x), Dv^\alpha)\, dy.$$

For $p: T \times W^{m,2}(W) \times W^{m,2}(W) \to R$ we assume that $p(t, x, x) \geqq c' \|x\|^p_{m,p}$ for all $x \in W^{m,2}(W)$. Let $A(t): W_0^{m,2}(W) \to W^{-m,2}(W) = (W_0^{m,2}(W))^*$ be defined by $p(t, x, v) = \langle A(t)x, v \rangle$ for all $x, v \in W_0^{m,2}(W)$. It is easy to see that because of (4), $A(t)(\cdot)$ is monotone.

We will show that $A(t)(\cdot)$ is sequentially weakly continuous. Let $x_n \xrightarrow{w} x$ in $W_0^{m,2}(W)$. Since $W_0^{m,2}(W) \hookrightarrow W_0^{m-1,2}(W)$ compactly (see Adams [1]), we have that $x_n \xrightarrow{s} x$ in $W_0^{m-1,2}(W)$. So if $\hat{A}_\alpha$ is the Nemitsky operator corresponding to A_α, from Krasnoselski's theorem we have that $\hat{A}_\alpha(t)(x_n) \to \hat{A}_\alpha(t)(x)$ in $L^2(W)$ and by passing to a subsequence if necessary, we may assume that $\hat{A}_\alpha(t)(x_n)(y) \to \hat{A}_\alpha(t)(x)(y)$ almost everywhere $\Rightarrow A_\alpha(t, y, x_n, Dx_n, \cdots, D^{m-1}x_n) \to A_\alpha(t, y, x, Dx, \cdots, D^{m-1}x)$ almost everywhere. Using (1) and (3) and the dominated convergence theorem, for every $z \in W_0^{m,2}(W)$ we get

$$\lim \langle A(t)x_n, z \rangle = \langle A(t)x, z \rangle$$

$$\Rightarrow A(t)x_n \xrightarrow{w} A(t)x \quad \text{in } W^{-m,2}(W),$$

and so we have shown the weak continuity of $A(t)(\cdot)$ on $W_0^{m,2}(W)$.

Also, through Pettis' Measurability Theorem (see Diestel and Uhl [18]), it is easy to check that $t \to A(t)x$ is measurable.

Furthermore,

$$\|A(t)x\|_{-m,2} \leqq \hat{g}(t) + c\|x\|_{m,2} \quad \text{a.e.}$$

with $\hat{g}(\cdot) \in L^2$ and from (5) we have that

$$\langle A(t)x, x \rangle \geqq c' \|x\|^2_{m,2}$$

for all $x \in W_0^{m,2}(W)$.

Let $f: T \times L^2(W) \to L^2(W)$ be defined by $f(t, u)(\cdot) = B(t)u(\cdot)$ and $Z = S^1_V$, which is w-compact in $L^2(T, L^2(W))$, and hence a compact Polish space for the weak topology.

Let $x_0(\cdot) \in W_0^{m,2}(W)$. Take $X = W_0^{m,2}(W)$, $H = L^2(W)$ and $X^* = W^{-m,2}(W)$. Then $(***)_2$ can be written as the following evolution equation:

$$(***)_2' \qquad \begin{aligned} &\dot{x}(t) + A(t, x(t)) = f(t, u(t)), \\ &x(0) = x_0, \qquad u \in Z. \end{aligned}$$

Then trajectories of this system are dense in the trajectories of the convexified system. Furthermore, those relaxed trajectories form a sequentially compact subset of $C(T, X_w)$.

Acknowledgment. The author expresses his gratitude to Professor L. Berkovitz and the referee for the constructive criticisms and suggestions that improved the material of this paper considerably.

REFERENCES

[1] R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] N. AHMED, *Properties of relaxed trajectories for a class of nonlinear evolution equations on a Banach space*, SIAM J. Control Optim., 21 (1983), pp. 953-967.
[3] N. AHMED AND K. TEO, *Optimal control of systems governed by a class of nonlinear evolution equation in a reflexive Banach space*, J. Optim. Theory Appl., 25 (1978), pp. 57-81.
[4] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, New York, 1984.
[5] E. BALDER, *A general denseness result for relaxed control theory*, Bull Austral. Math. Soc., 30 (1984), pp. 463-475.
[6] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.
[7] L. BERKOVITZ, *Optimal Control Theory*, Appl. Math. Sci., 12 (1983).
[8] J. BROOKS AND N. DINCULEANU, *Conditional expectations and strong compactness in spaces of Bahner integrable functions*, J. Multivariate Anal., 9 (1979), pp. 420-427.
[9] R. CARROLL, *Abstract Methods in Partial Differential Equations*, Harper and Row, New York, 1969.
[10] C. CASTAING, *Quelques apercus des resultats de compacité dans* L_E^p $(1 \leqq p < \infty)$, Sém. Anal. Convexe, 16 (1980).
[11] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 580, Springer-Verlag, Berlin, New York, 1977.
[12a] L. CESARI, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints* I, Trans. Amer. Math. Soc., 124 (1966), pp. 369-412.
[12b] ———, *Existence theorems for weak and usual optimal solutions in Lagrange problems with unilateral constraints* II, Trans. Amer. Math. Soc., 124 (1966), pp. 413-427.
[13] ———, *Optimization: Theory and Applications*, Appl. Math., 17 (1983).
[14] G. CHOQUET, *Lectures on Analysis*, Vol. 1, W. A. Benjamin, London, 1969.
[15] P. V. CHUONG, *On the density of extremal selections for measurable multifunctions*, Acta Math. Vietnam., 6 (1981), pp. 13-28.
[16] C. DELLACHERIE, *Ensembles analytiques: Théorèmes de séparation et applications*, Séminaires de Probabilités IX, Lecture Notes in Math. 465, Springer-Verlag, Berlin, New York, 1975.
[17] C. DELLACHERIE AND P.-A. MEYER, *Probabilities and Potential*, North-Holland, Amsterdam, New York, 1978.
[18] J. DIESTEL AND J. UHL, *Vector Measures*, Math. Surveys Monogr. 15, American Mathematical Society, Providence, RI, 1977.
[19] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Vol. I, John Wiley, New York, 1958.
[20] G. EDGAR, *Measurability in a Banach space* II, Indiana Univ. Math. J., 28 (1979), pp. 559-578.
[21] A. FILIPPOV, *On certain questions in the theory of optimal control*, SIAM J. Control Ser. A, 1 (1962), pp. 76-84.
[22] R. GAMKRELIDZE, *Principles of Optimal Control Theory*, Plenum Press, New York, 1978.
[23] S. GUTMAN, *Compact perturbations of m-accretive operators in general Banach spaces*, SIAM J. Math. Anal., 13 (1982), pp. 789-800.
[24] H. HERMES AND J. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[25] F. HIAI AND H. UMEGAKI, *Integrals, conditional expectations and martingales of multivalued functions*, J. Multivariate Anal., 7 (1977) pp. 149-182.
[26] A. JAWHAR, *Mesures de transition et applications*, Sém. Anal. Convexe, 13 (1984).
[27] D. KANDILAKIS AND N. S. PAPAGEORGIOU, *On the properties of the Aumann integral with applications to differential inclusions and control systems*, Czechoslovak Math. J., to appear.
[28] V. LAKSHMIKANTHAM AND S. LEELA, *Nonlinear Differential Equations in Abstract Spaces*, Pergamon, Oxford, 1981.
[29] R. MARTIN, *Nonlinear Operators and Differential Equations in Banach Spaces*, John Wiley, New York, 1976.
[30] N. S. PAPAGEORGIOU, *On multivalued evolution equations and differential inclusions in Banach spaces*, Comment. Math. Univ. St. Paul., 36 (1987), pp. 21-39.
[31] ———, *On the theory of Banach space valued multifunctions. Part* 1: *Integration and conditional expectation*, J. Multivariate Anal., 17 (1985), pp. 185-207.
[32] N. PAVEL, *Differential Equations, Flow Invariance and Applications*, Research Notes in Math. 113, Pitman, Boston, 1984.
[33] M.-F. SAINT-BEUVE, *On the extension of Von Neumann-Aumann's theorem*, J. Funct. Anal., 17 (1974), pp. 112-129.
[34] ———, *Une extension des théorèmes de Novikov et d'Arsenin*, Sém. Anal. Convexe, 18 (1981).
[35] H. TANABE, *Equations of Evolution*, Pitman, London, 1979.
[36] D. WAGNER, *Survey of measurable selection theorems*, SIAM J. Control Optim., 15 (1977), pp. 859-903.
[37] J. WARGA, *Relaxed variational problems*, J. Math. Anal. Appl., 4 (1962), pp. 111-128.
[38] ———, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[39] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 289-295, March 1989

003

EXISTENCE OF OPTIMAL STRATEGIES IN ZERO-SUM NONSTATIONARY STOCHASTIC GAMES WITH LACK OF INFORMATION ON BOTH SIDES*

ANDRZEJ S. NOWAK†

Abstract. This paper studies a zero-sum discrete-time stochastic game model with Borel state and action spaces. The law of motion of the system in the model is assumed to be nonstationary. Following M. Schäl, at each stage of the game every player is assumed to know the sequence of states occurring up to this stage, but has no explicit information about his opponent's previous decisions. Under certain semicontinuity and compactness conditions, the existence of a value is proved for such a game and the existence of optimal (ε-optimal) universally measurable strategies for the minimizer (maximizer). This essentially improves a result of Schäl on this subject.

Key words. zero-sum nonstationary stochastic games, imperfect information, minimax theorem, universally measurable strategies

AMS(MOS) subject classifications. primary 90D15; secondary 60K99, 93C55

1. Introduction and a minimax theorem. In this paper we study a two-person zero-sum nonstationary stochastic game introduced by Schäl [18]. It is assumed that at every stage of the game neither player has any explicit information about his opponent's earlier choices. We assume, however, that every player knows the sequence of states of the system occurring up to this stage and remembers all his previous choices. Our aim is to prove that in Schäl's game every player has an optimal (or ε-optimal, $\varepsilon > 0$) strategy independent of any probability distribution of the initial state. This is an essential improvement of Schäl's theorem from [18]. For a more extensive discussion of our results, see Remarks 1-4.

Let X be a Borel space, i.e., a nonempty Borel subset of a complete separable metric space. We assume that X is endowed with the relative topology and the Borel σ-algebra $\mathcal{B}(X)$. By $\mathcal{U}(X)$ we denote the σ-algebra of all universally measurable subsets of X (see [1, Chap. 7, Appendix B]). Clearly, $\mathcal{B}(X) \subset \mathcal{U}(X)$. Let $P(X)$ be the space of all probability measures on $\mathcal{B}(X)$. We shall assume that for every Borel space X, $P(X)$ is given the weak topology (see [1, Chap. 7]).

Let X and Y be Borel spaces. By a universally measurable transition probability from X to Y we mean a universally measurable mapping from X to $P(Y)$. It is known that every universally measurable transition probability $f: X \to P(Y)$ may be identified with a function $f(\cdot|\cdot)$, so that, for each $x \in X$, $f(\cdot|x) \in P(Y)$, and for each $B \in \mathcal{B}(Y)$, $f(B|\cdot)$ is a universally measurable mapping from X to $[0,1]$ (see [1, Lemma 7.28]).

Let $X_1, X_2, \cdots$ be a sequence of nonempty sets. The Cartesian products of $X_1, \cdots, X_n$ and $X_1, X_2, \cdots$, are denoted by $X_1 \cdots X_n$ and $X_1 X_2 \cdots$, respectively. Let $X_1, X_2, \cdots$ be Borel spaces. Throughout this paper we assume that the product spaces $X_1 \cdots X_n$ and $X_1 X_2 \cdots$ are given their product topologies and product σ-algebras. It is well known that the product σ-algebra $\mathcal{B}(X_1) \cdots \mathcal{B}(X_n)$ in $X_1 \cdots X_n$ is equal to $\mathcal{B}(X_1 \cdots X_n)$. A similar result is also valid for the product space $X_1 X_2 \cdots$, [1, Prop. 7.13].

* Received by the editors December 15, 1986; accepted for publication (in revised form) May 17, 1988.
† Institute of Mathematics, Wrocław Technical University, Wyspiańskiego 27, 50-370 Wrocław, Poland.

A *zero-sum nonstationary stochastic game with lack of information on both sides* is defined by a sequence of objects $\{S_n, A_n, B_n, t_{n+1}, u_n;\ n \in N\}$, where:

(i) S_n is the *state space* at time $n \in N$ and is assumed to be a Borel space.

(ii) A_n and B_n are the *action spaces* at time $n \in N$ of players I and II, respectively. It is assumed that A_n is a Borel space and B_n is a nonempty compact metric space.

Let $H_1 = S_1$, $H_n = S_1 A_1 B_1 \cdots S_{n-1} A_{n-1} B_{n-1} S_n$, $H_\infty = S_1 A_1 B_1 S_2 A_2 B_2 \cdots$. Then H_n is the set of histories of the game for horizon $n \in N$, while H_∞ is the set of all infinite histories of the game.

(iii) $\{t_{n+1}\}$ is the *law of motion* of the system; t_{n+1} is a Borel measurable transition probability from $H_n A_n B_n$ to S_{n+1}, $n \in N$. We assume that t_{n+1} is dominated by a probability measure $q_{n+1} \in P(S_{n+1})$. The density c_{n+1} of t_{n+1} with respect to q_{n+1} is assumed to be a Borel measurable function on H_{n+1} such that, for each $a_1, \cdots, a_n$ and $s_1, \cdots, s_n$, s_{n+1}, $c_{n+1}(s_1, a_1, \cdot, \cdots, s_n, a_n, \cdot, s_{n+1})$ is lower semicontinuous on $B_1 \cdots B_n$.

(iv) u_n is the *payoff function* of player I at stage $n \in N$ ($-u_n$ is the payoff function for player II). It is assumed that u_n is a nonnegative Borel measurable function on $H_n A_n B_n$ such that, for each $s_1, \cdots, s_n$ and $a_1, \cdots, a_n$, $u_n(s_1, a_1, \cdot, \cdots, s_n, a_n, \cdot)$ is lower semicontinuous on $B_1 \cdots B_n$. Of course, every u_n may be recognized as a function on H_∞. Doing so, we assume that the sequence $\{u_n\}$ is nondecreasing and $u = \lim_n u_n$ is the *infinite horizon payoff function* for player I.

The game is played as follows. The players observe an initial state $s_1 \in S_1$ and independently choose actions $a_1 \in A_1$ and $b_1 \in B_1$, respectively. Then the system moves to a new state $s_2 \in S_2$ according to the probability distribution $t_2(\cdot \mid s_1, a_1, b_1)$ on which, knowing $s_2 \in S_2$, the players choose independently $a_2 \in A_2$ and $b_2 \in B_2$, respectively. The system moves to a new state $s_3 \in S_3$ according to the probability distribution $t_3(\cdot \mid s_1, a_1, b_1, s_2, a_2, b_2)$ and so on. The result of such an infinite sequence of moves is a point $h \in H_\infty$ and player II pays player I the amount $u(h)$. Note that at each stage n of the game every player knows the sequence of states $s_1, \cdots, s_n$ and his own previous choices.

We put $H_1^1 = H_1^2 = S_1$, $H_n^1 = S_1 A_1 \cdots S_{n-1} A_{n-1} S_n$, $H_n^2 = S_1 B_1 \cdots S_{n-1} B_{n-1} S_n$, for $n \geqq 2$. Let $F_n(G_n)$ be the set of all universally measurable transition probabilities from $H_n^1(H_n^2)$ to $A_n(B_n)$. A universally measurable *strategy* for player I(II) is a sequence $f = \{f_n\}$ ($g = \{g_n\}$), where $f_n \in F_n(g_n \in G_n)$ for each $n \in N$. Denote by $F(G)$ the set of all strategies for player I(II).

Let $E_{f_n}, E_{g_n}, E_{t_{n+1}}$ denote the conditional expectation operator with respect to $f_n \in F_n$, $g_n \in G_n$, t_{n+1}, respectively. By the Ionescu-Tulcea Theorem [10, Prop. V.1.1] and [1, Lemma 7.28], each pair of strategies $f = \{f_n\}$, $g = \{g_n\}$, together with the law of motion $\{t_{n+1}\}$, uniquely defines a universally measurable transition probability $P_{fg}(\cdot \mid \cdot)$ from S_1 to $A_1 B_1 S_2 A_2 B_2 S_3 \cdots$ such that, for every Borel measurable function w: $H_n A_n B_n \to R$ ($n \in N$) bounded below, we have

$$E(w, f, g)(s_1) := \int w(h) P_{fg}(dh \mid s_1) = (E_{f_1} E_{g_1} E_{t_2} \cdots E_{f_{n-1}} E_{g_{n-1}} E_{t_n} E_{f_n} E_{g_n} w)(s_1),$$

where $s_1 \in S_1$. (Here w is also regarded as a function on H_∞.) Thus, each pair f, g of strategies defines an *expected payoff* to player I at an initial state $s_1 \in S_1$ to be

$$E(u, f, g)(s_1) = \int u(h) P_{fg}(dh \mid s_1).$$

Under our assumption (iv), from the monotone convergence theorem and Fubini's theorem we infer that, for each $s_1 \in S_1$, $f = \{f_n\} \in F$, $g = \{g_n\} \in G$,

$$
\begin{aligned}
E(u,f,g)(s_1) &= \lim_n E(u_n,f,g)(s_1) \\
&= \lim_n (E_{f_1}E_{g_1}E_{t_2}\cdots E_{f_{n-1}}E_{g_{n-1}}E_{t_n}E_{f_n}E_{g_n}u_n)(s_1) \\
&= \lim_n (E_{g_1}E_{f_1}E_{t_2}\cdots E_{g_{n-1}}E_{f_{n-1}}E_{t_n}E_{g_n}E_{f_n}u_n)(s_1).
\end{aligned} \tag{1}
$$

From now on we assume that

(v) $\quad E(u,f,g)(s_1) < \infty$ for all $s_1 \in S_1, f \in F, g \in G$.

Define, for each initial state $s_1 \in S_1$,

$$
v_*(s_1) = \sup_{f\in F}\inf_{g\in G} E(u,f,g)(s_1) \quad \text{and} \quad v^*(s_1) = \inf_{g\in G}\sup_{f\in F} E(u,f,g)(s_1).
$$

Then $v_*(v^*)$ is called the *lower-* (*upper-*) *value function* of the game. If $v_* = v^*$, then this common function is called the *value function* of the game and is denoted by v.

Suppose the value function v exists. A strategy $f^* \in F$ is called *ε-optimal* for player I for given $\varepsilon > 0$ if

$$
\inf_{g\in G} E(u,f^*,g)(s_1) + \varepsilon > v(s_1) \quad \text{provided that } v(s_1) < \infty,
$$

$$
\inf_{g\in G} E(u,f^*,g)(s_1) > \frac{1}{\varepsilon} \quad \text{if } v(s_1) = \infty.
$$

A strategy $g^* \in G$ is called *optimal* for player II if

$$
\sup_{f\in F} E(u,f,g^*)(s_1) \leqq v(s_1) \quad \text{for all } s_1 \in S_1.
$$

Here is the main result of this paper.

MINIMAX THEOREM. *Assume* (i)-(v). *Then the stochastic game has a value function, which is universally measurable. Player* II *has an optimal universally measurable strategy while, for any* $\varepsilon > 0$, *player* I *has an* ε*-optimal universally measurable strategy.*

Remarks. (1) The same information structure is considered in Schäl's stochastic game model [18]. It should be noted, however, that, for any pair of strategies $f \in F$ and $g \in G$, the payoff for player I in Schäl's game is the expectation of $E(u,f,g)(\cdot)$ with respect to a probability measure $p \in P(S_1)$, called the probability distribution of the initial state. Thus, the optimal (ε-optimal) strategies in Schäl's approach depend on p. Such strategies are often called in the literature $\bar{p}$-optimal (or ε-$\bar{p}$-optimal) (see [6], [16], [18]). The optimality criterion considered in this paper is essentially stronger than that of Schäl [18]. For example, every optimal strategy for player II is $\bar{p}$-optimal for each probability distribution p of the initial state.

(2) From [18], it follows that our game has a value for each initial state, but to prove the remaining details of the Minimax Theorem we have to do some extra work. We consider an auxiliary game in which the strategy sets are the spaces $\Pi(A)$ and $\Pi(B)$ of probability measures induced by strategies of players I and II, respectively (see § 2). Our approach is then based on the equivalence of the weak topology and the ws^∞-topology of Schäl [16] in $\Pi(B)$ proved by Nowak in [15] (see Lemma 6) and the minimax selection theorem of [12]. We note that the issues described above do not arise, if we restrict attention only to $\bar{p}$-optimal strategies as in [18].

(3) Our convergence assumption (iv) includes the discounted and positive stochastic games (see, for example, [9], [11] and the references therein). It is worth noting

that optimal strategies for player I need not exist even if he has finite action spaces (see, for example, [3], [9]). Therefore, we make no compactness assumptions regarding the action spaces of player I.

(4) Imposing certain semicontinuity and compactness conditions on player I, as well as a stronger convergence assumption on the n-stage payoff functions u_n, we can restrict ourselves to Borel measurable strategies (see [12, Remark 3]).

(5) Universally measurable strategies have been broadly used in control and gambling theory (see, for example, [1], [2], [7]). They have been applied to zero-sum stochastic games by Nowak in [11], [12], and [14].

(6) Nonstationary stochastic games with standard information structures where, at each stage of the play, every player has full information about the states of the system and decisions made by the players in the past, have already been studied in [13], [14], and some of the references therein. It should be noted that the approaches taken in [13] and [14] cannot be applied to the present model. We hope that the methods used in this paper can also be applied to some more general models of games of incomplete information.

2. Proof of the main result. Let $\bar{F}_n(\bar{G}_n)$ be the set of all transition probabilities from $F_n(G_n)$ that are independent of the initial state $s_1 \in S_1$. We put $\bar{F} = \bar{F}_1\bar{F}_2 \cdots$ and $\bar{G}_1\bar{G}_2 \cdots$, and further, we put $A = A_1S_2A_2S_3 \cdots$ and $B = B_1S_2B_2S_3 \cdots$. To each $f = \{f_n\} \in \bar{F}$ is associated a probability measure $P_f = f_1q_2f_2q_3 \cdots$ on $\mathcal{B}(A)$, given by the product measure theorem of Ionescu Tulcea (see [1, Prop. 7.45] or [10, Prop. V.1.1]). (Recall our assumption (iii).) Similarly, to each $g \in \bar{G}$ is associated a probability measure P_g on $\mathcal{B}(B)$.

We put $\Pi(A) = \{P_f \in P(A): f \in \bar{F}\}$ and $\Pi(B) = \{P_g \in P(B): g \in \bar{G}\}$.

From the proof of Proposition 7.45 in [1], we infer the following lemma.

LEMMA 1. *For each $f \in \bar{F}$ and $g \in \bar{G}$, there are Borel measurable transition probabilities $f_n \in \bar{F}_n$ and $g_n \in \bar{G}_n$, $n \in N$ so that $P_f = P_{\bar{f}}$ and $P_g = P_{\bar{g}}$, where $\bar{f} = \{\bar{f}_n\}$ and $\bar{g} = \{\bar{g}_n\}$.*

In the remainder of this paper, we assume that $P(A)$, $P(B)$ are given the weak topologies and $\Pi(A)$, $\Pi(B)$ are endowed with the relative topologies and Borel σ-algebras.

LEMMA 2. (a) $\Pi(A)$ *is a convex Borel subset of a Borel space $P(A)$.*

(b) $\Pi(B)$ *is a compact convex subset of a Borel space $P(B)$.*

Proof. By Lemma 1, we can restrict ourselves to Borel measurable transition probabilities in the definitions of $\Pi(A)$ and $\Pi(B)$. The spaces $P(A)$ and $P(B)$ are Borel by [1, Prop. 7.13 and Cor. 7.25.1]. Now (a) follows from [17, Thm. 7.11] and [19, Lemma 7.2] while (b) is a corollary to [16, Thm. 5.6] and [17, Thm. 7.11]. □

Results closely related to Lemma 3 below have appeared in dynamic programming and gambling literature (see [19, p. 885], [8, § 4], [7, Lemma 3.3(b), Remark 3.1]).

LEMMA 3. (a) *Let $f = \{f_n\}$ be any strategy of player* I. *Then $s_1 \to (f_1q_2f_2q_3 \cdots)(\cdot|s_1)$ is a universally measurable mapping from S_1 to $\Pi(A)$.*

(b) *Let $\varphi: S_1 \to \Pi(A)$ be a universally measurable mapping. Then there exist universally measurable transition probabilities $f_n \in F_n$, $n \in N$, so that $\varphi(s_1)(\cdot) = (f_1q_2f_2q_3 \cdots)(\cdot|s_1)$, for every $s_1 \in S_1$.*

Proof. Part (a) follows from [10, Prop. V.1.1] and [1, Lemma 7.28]. The proof of (b) is similar to that of Lemma 3.3(b) in [7]. The transition probabilities in (b) are constructed inductively by means of [1, Prop. 7.27] and a characterization of the space of probability measures induced by strategies (alias policies) given by Strauch [19, Lemma 7.2] and Hinderer [6, Lemma 13.1]. □

Remark 7. Replacing $\Pi(A)$ by $\Pi(B)$ in Lemma 3, we get a similar result concerning strategies of player II.

Let us put $A^n = A_1 \cdots A_n$, $B^n = B_1 \cdots B_n$ and $S^n = S_2 \cdots S_n$, $n \geq 2$. The elements of A^n, B^n, and S^n will be denoted by a^n, b^n, and s^n, respectively.

For each $f = \{f_n\} \in \bar{F}$ and $n \geq 2$, there exists a universally measurable transition probability $f^n = f_1 \cdots f_n$ from S^n to A^n such that, for every bounded below Borel measurable function w: $S^n A^n \to R$, we have

$$(E_{f^n} w)(s^n) := \int w(s^n, a^n) f^n(da^n \mid s^n) = (E_{f_1} \cdots E_{f_n} w)(s^n), \qquad s^n \in S^n,$$

(see [1, pp. 175-177] or [10, Prop. V.1.1]). Here $E_{f^n} w$ is the conditional expectation of w with respect to f^n. Similarly, to each $g = \{g_n\} \in \bar{G}$, we associate universally measurable transition probabilities $g^n (n \geq 2)$ from S^n to B^n. We write E_{g^n} for the conditional expectation with respect to g^n.

Let q^n be the product of probability measures $q_2, \cdots, q_n$. The expectation with respect to q^n is denoted by E_{q^n}.

LEMMA 4. *Assume $P_f = P_{\bar{f}}$ for some $f = \{f_n\}$ and $\bar{f} = \{\bar{f}_n\}$ belonging to $\bar{F}$. Then, for each $n \geq 2$, we have $f^n(\cdot \mid s^n) = \bar{f}^n(\cdot \mid s^n)$ almost everywhere q^n.*

Proof. Let $n \geq 2$ be fixed. From [1, Props. 7.19, 7.20], it follows that there exists a sequence $\{w_m\}$ of bounded continuous functions on A^n that separate elements of $P(A^n)$. Of course, for each $m \in N$ and every Borel subset C of S^n, we have

$$\int w_m \chi_C \, dP_f = \int w_m \chi_C \, dP_{\bar{f}},$$

where χ_C means the characteristic function of C. By Fubini's theorem, it follows that $E_{q^n} E_{f^n} w_m \chi_C = E_{q^n} E_{\bar{f}^n} w_m \chi_C$, for each $C \in \mathcal{B}(S^n)$, $m \in N$. Thus, for every $m \in N$, there is $C_m \in \mathcal{B}(S^n)$ with $q^n(C_m) = 0$ and $E_{f^n} w_m = E_{\bar{f}^n} w_m$ on $S^n - C_m$. Define $C = \bigcup_{m \in N} C_m$. Then $q^n(C) = 0$ and $E_{f^n} w_m = E_{\bar{f}^n} w_m$ on $S^n - C$, for each $m \in N$. Since $\{w_m\}$ separates elements of $P(A^n)$, the result follows. □

Remark 8. It is obvious that if $P_g = P_{\bar{g}}$ for some $g, \bar{g} \in \bar{G}$, then a similar result is in force for the transition probabilities g^n and $\bar{g}^n (n \geq 2)$ determined by g and $\bar{g}$, respectively.

Let u_n: $H_n A_n B_n \to R$ be an n-stage payoff function of player I. From our assumption (iii) and (1), it follows that, for each $s_1 \in S_1$, $f = \{f_n\} \in \bar{F}$, and $g = \{g_n\} \in \bar{G}$, we have (cf. [17, § 7])

$$E(u_n, f, g)(s_1) = (E_{q^n} E_{f^n} E_{g^n} u_n c_2 \cdots c_n)(s_1) = (E_{q^n} E_{g^n} E_{f^n} u_n c_2 \cdots c_n)(s_1). \tag{2}$$

Moreover, we have

$$E(u_n, f, g)(s_1) = (E_\varphi E_{g^n} u_n c_2 \cdots c_n)(s_1) = (E_\gamma E_{f^n} u_n c_2 \cdots c_n)(s_1), \tag{3}$$

where $E_\varphi (E_\gamma)$ is the expectation with respect to $\varphi = P_f$ ($\gamma = P_g$), $s_1 \in S_1$.

Define $F_\varphi = \{f \in \bar{F}: \varphi = P_f\}$, $G_\gamma = \{g \in \bar{G}: \gamma = P_g\}$, where $\varphi \in \Pi(A)$, $\gamma \in \Pi(B)$. From Lemma 4, Remark 8, and (2) it follows that if $\varphi \in \Pi(A)$ and $\gamma \in \Pi(B)$, then $E(u_n, f, g)(s_1)$ has the same value for every $f \in F_\varphi$ and $g \in G_\gamma$. Therefore, for each $\varphi \in \Pi(A)$ and $\gamma \in \Pi(B)$ and $s_1 \in S_1$, we may define

$$U_n(\varphi, \gamma)(s_1) := E(u_n, f, g)(s_1),$$

where $f(g)$ is an arbitrary element of $F_\varphi (G_\gamma)$. Further, for every $\varphi \in \Pi(A)$, $\gamma \in \Pi(B)$, and $s_1 \in S_1$, we put

$$U(\varphi, \gamma)(s_1) = \lim_n U_n(\varphi, \gamma)(s_1).$$

By the monotone convergence theorem

$$U(\varphi, \gamma)(s_1) = E(u, f, g)(s_1), \tag{4}$$

for every $s_1 \in S_1$, $\varphi \in \Pi(A)$, $\gamma \in \Pi(B)$, and $f \in F_\varphi$, $g \in G_\gamma$.

The following lemma is obvious.

LEMMA 5. *For each initial state $s_1 \in S_1$, we have*

$$v_*(s_1) = \sup_{f \in \bar{F}} \inf_{g \in \bar{G}} E(u, f, g)(s_1) = \sup_{\varphi \in \Pi(A)} \inf_{\gamma \in \Pi(B)} U(\varphi, \gamma)(s_1),$$

$$v^*(s_1) = \inf_{g \in \bar{G}} \sup_{f \in \bar{F}} E(u, f, g)(s_1) = \inf_{\gamma \in \Pi(B)} \sup_{\varphi \in \Pi(A)} U(\varphi, \gamma)(s_1).$$

Recall that $\Pi(A)$ ($\Pi(B)$) is given the relative weak topology from $P(A)(P(B))$ and the Borel σ-algebra.

LEMMA 6. *The function $U : \Pi(A)\Pi(B)S_1 \to R$ defined by* (4) *is Borel measurable. Moreover, for each $s_1 \in S_1$, $\varphi \in \Pi(A)$, the function $U(\varphi, \cdot)(s_1)$ is lower semicontinuous on $\Pi(B)$.*

Proof. Let $n \geqq 2$ be fixed. The function $u_n c_2 \cdots c_n$ is Borel measurable and lower semicontinuous in the actions of player II. Under our compactness assumption (ii), there exists a nondecreasing sequence of bounded Borel measurable functions $w_k : H_n A_n B_n \to R$, $k \in N$, such that $w_k \uparrow u_n c_2 \cdots c_n$ as $k \to \infty$, and, for each $k \in N$, w_k is continuous in the actions of player II [17, Eq. (4.1)].

Define

$$W_k(\varphi, \gamma)(s_1) := (E_\gamma E_{f^n} w_k)(s_1), \qquad k \in N,$$

where $s_1 \in S_1$, $\varphi \in \Pi(A)$, $\gamma \in \Pi(B)$, and f is an arbitrary element of F_φ. Clearly, W_k is a well-defined function on $\Pi(A)\Pi(B)S_1$. Note that, for each $\varphi \in \Pi(A)$, $f \in F_\varphi$, $E_{f^n} w_k$ is a bounded Borel measurable function on $S_1 S^n B^n$. Moreover, $E_{f^n} w_k$ depends continuously with respect to $b^n \in B^n$. By Lemma 1 and [15, Thm. 1, Remark 2], the function $\gamma \to (E_\gamma E_{f^n} w_k)(s_1)$ is continuous on $\Pi(B)$ for each $\varphi \in \Pi(A)$, $f \in F_\varphi$, $s_1 \in S_1$. (This is a consequence of the equivalence of the weak topology on $\Pi(B)$ and the ws^∞-topology of Schäl [16] proved in [15].)

Let $\gamma \in \Pi(B)$ be fixed. Then $W_k(\varphi, \gamma)(s_1) = (E_\varphi E_{g^n} w_k)(s_1)$, $s_1 \in S_1$, $\varphi \in \Pi(A)$, and g is an arbitrary element of G_γ. Applying standard arguments and Proposition 7.25 of [1], we infer that $(\varphi, s_1) \to (E_\varphi E_{g^n} w_k)(s_1)$ is a Borel measurable function on $\Pi(A)S_1$, for each $g \in G_\gamma$. Thus, we have shown that $W_k(\varphi, \cdot)(s_1)$ is continuous on $\Pi(B)$, for every $s_1 \in S_1$, $\varphi \in \Pi(A)$, $k \in N$, and $W_k(\cdot, \gamma)(\cdot)$ is Borel measurable on $\Pi(A)S_1$, for each $\gamma \in \Pi(B)$, $k \in N$. This implies that W_k is Borel measurable [5, Thm. 6.1]. By the monotone convergence theorem, $W_k \uparrow U_n$ on $\Pi(A)\Pi(B)S_1$ as $k \to \infty$, and $U_n \uparrow U$ as $n \to \infty$. This completes the proof of the required properties of the function U. □

LEMMA 7. *For each $s_1 \in S_1$, $\varphi \in \Pi(A)$, and $\gamma \in \Pi(B)$ the functions $U(\cdot, \gamma)(s_1)$ and $U(\varphi, \cdot)(s_1)$ are affine on $\Pi(A)$ and $\Pi(B)$, respectively.*

Proof. The proof follows from (3) and Lemma 2. □

Proof of Minimax Theorem. By Fan's Minimax Theorem [4, Thm. 2] and Lemmas 2, 6, and 7, the zero-sum game with payoff function $U(\cdot, \cdot)(s_1)$ has a value, for each $s_1 \in S_1$, which (by Lemma 5) is equal to $v(s_1)$, the value of our stochastic game at an initial state s_1. Using Lemmas 2, 6, and the minimax selection theorem given as Theorem 2.2 of [12], we infer that the value function is universally measurable. Moreover, from

[12, Thm. 2.2], it follows that there exists a universally measurable mapping $\gamma^*: S_1 \to \Pi(B)$ such that

$$\sup_{\varphi \in \Pi(A)} U(\varphi, \gamma^*(s_1))(s_1) = v(s_1) \quad \text{for all } s_1 \in S_1,$$

and, for each $\varepsilon > 0$, there exists a universally measurable mapping $\varphi^*: S_1 \to \Pi(A)$ such that

$$\inf_{\gamma \in \Pi(B)} U(\varphi^*(s_1), \gamma)(s_1) + \varepsilon > v(s_1) \quad \text{provided that } v(s_1) < \infty,$$

$$\inf_{\gamma \in \Pi(B)} U(\varphi^*(s_1), \gamma)(s_1) > \frac{1}{\varepsilon} \quad \text{when } v(s_1) = \infty.$$

Appealing now to Lemma 3, Remark 7, (3), and (4) we note that $\gamma^*(\varphi^*)$ determines an optimal (ε-optimal) strategy for player II (player I). □

Acknowledgments. The author is grateful to the editor and two anonymous referees for splendid editorial work.

REFERENCES

[1] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[2] D. BLACKWELL, D. FREEDMAN, AND M. ORKIN, *The optimal reward operator in dynamic programming*, Ann. Probab., 2 (1974), pp. 926-941.

[3] H. EVERETT, *Recursive games*, Ann. of Math. Stud., 39 (1957), pp. 47-78.

[4] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. U.S.A., 39 (1953), pp. 42-47.

[5] C. J. HIMMELBERG, *Measurable relations*, Fund. Math., 87 (1975), pp. 53-72.

[6] K. HINDERER, *Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.

[7] R. P. KERTZ, *Renewal plans and persistent optimality in countably additive gambling*, Math. Oper. Res., 7 (1982), pp. 361-382.

[8] R. P. KERTZ AND D. NACHMAN, *Persistently optimal plans for nonstationary dynamic programming*, Ann. Probab., 7 (1979), pp. 811-826.

[9] P. R. KUMAR AND T. H. SHIAU, *Existence of value and randomized strategies in zero-sum discrete-time stochastic dynamic games*, SIAM J. Control Optim., 19 (1981), pp. 617-634.

[10] J. NEVEU, *Mathematical Foundations of the Calculus of Probability*, Holden-Day, San Francisco, 1965.

[11] A. S. NOWAK, *Universally measurable strategies in zero-sum stochastic games*, Ann. Probab., 13 (1985), pp. 269-287.

[12] ———, *Measurable selection theorems for minimax stochastic optimization problems*, SIAM J. Control Optim., 23 (1985), pp. 466-476.

[13] ———, *Semicontinuous nonstationary stochastic games*, J. Math. Anal. Appl., 117 (1986), pp. 84-99.

[14] ———, *Semicontinuous nonstationary stochastic games*, II, J. Math. Anal. Appl., to appear.

[15] ———, *On the weak topology on a space of probability measures induced by policies*, Bull. Polish Acad. Sci., Ser. Math., to appear.

[16] M. SCHÄL, *On dynamic programming: compactness of the space of policies*, Stochastic Process. Appl., 3 (1975), pp. 345-364.

[17] ———, *On dynamic programming and statistical decision theory*, Ann. Statist., 7 (1979), pp. 432-445.

[18] ———, *Stochastic nonstationary two-person zero-sum games*, Z. Angew. Math. Mech., 61 (1981), pp. 352-353.

[19] R. E. STRAUCH, *Negative dynamic programming*, Ann. Math. Statist., 37 (1966), pp. 871-890.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 296-307, March 1989

004

THE EXISTENCE OF A MINIMUM PAIR OF STATE AND POLICY FOR MARKOV DECISION PROCESSES UNDER THE HYPOTHESIS OF DOEBLIN*

M. KURANO†

Abstract. This paper studies the average-cost Markov decision process with compact state and action spaces and bounded lower semicontinuous cost functions. Following the idea of Borkar's excellent papers [*SIAM J. Control Optim.*, 21 (1983), pp. 652-666; 22 (1984), pp. 965-978], the general case where irreducibility is not assumed is considered under the hypothesis of Doeblin and the existence of a minimum pair of state and policy, which attains the infimum of the average expected cost over all initial states and policies, is established. Further, it is proved that under additional weak conditions there exists an optimal stationary policy in the usual sense.

Key words. Markov decision process, average cost criterion, Doeblin condition

AMS(MOS) subject classifications. primary 90C40; secondary 90C39

1. Introduction and notation. The study of the average-cost Markov decision process with general state and action spaces has been done by Ross [14], Tijms [15], Kurano [10], and others. But as far as the author is aware, they have treated the case in which there is a single ergodic class and approximate recurrency under all policies.

In this paper, we treat with the general case, in which several ergodic classes and transient sets are permitted for the Markov process induced by any randomized stationary policy, under the compactness of state and action spaces. And introducing the hypothesis of Doeblin [7], we show the existence of a minimum pair of state and policy, which attains the infimum of the average expected cost over all initial states and policies. Further, under additional weak conditions it is proved that there exists an optimal stationary policy in the usual sense. Here we do not use the traditional approach taken by Ross [14], Tijms [15], and others which treats the problem as a limiting case of the discounted cost problem. To prove the existence of a minimum pair of state and policy we apply the direct method by the empirical distribution, which Borkar [4], [5] uses to establish the existence of optimal stationary policies for countable-state Markov decision processes in cases where a single communicating class exists under any stationary policy.

For examples of treatment of the general (multichain) case with the finite or denumerable state space, see Deppe [6], Howard [9], Schweitzer [16], and Zijm [17].

In the remainder of this section, we shall establish the notation that will be used throughout the paper and define the problem to be examined. Also, a minimum pair of state and policy is defined. In §§ 2 and 3, we give the existence and characterization of minimum pairs under the hypothesis of Doeblin. In § 2 it is proved that there exists a subset C of the state space such that there is a minimum pair only for all initial states belonging to C, implying the existence of an optimal stationary policy for these initial states. The extension of these results to all of the state space is done under additional assumptions in § 3. Finally, in § 4, the more general case is discussed under some continuity conditions.

A Borel set is a Borel subset of a complete separable metric space. For a Borel set X, $\mathcal{B}_X$ denotes the Borel subsets of X. A Markov decision process is a controlled

* Received by the editors July 27, 1987; accepted for publication (in revised form) May 18, 1988.
† Department of Mathematics, Faculty of Education, Chiba University, Yayoi-cho, Chiba, 260, Japan.

dynamic system defined by four objects: S, $\{A(x), x \in S\}$, c, and Q, where S is any Borel set representing the state space of some system and for each $x \in S$, the admissible action space $A(x)$ is a nonempty subset of some Borel set A such that $\{(x, a): x \in S, a \in A(x)\}$ is an element of $\mathcal{B}_S \times \mathcal{B}_A$, the immediate cost function c is a real-valued Borel measurable function on $S \times A$, and $Q(\cdot | x, a)$ is the law of motion, which is taken to be stochastic kernel on $\mathcal{B}_S \times S \times A$; i.e., for each $(x, a) \in S \times A$, $Q(\cdot | x, a)$ is a probability measure on $\mathcal{B}_S$; and, for each $D \in \mathcal{B}_S$, $Q(D|\cdot)$ is a Borel measurable function on $S \times A$.

Throughout this paper, the following assumptions will remain operative:

(i) S and $R := \{(x, a) | x \in S, a \in A(x)\}$ are compact;

(ii) c is a nonnegative real-valued bounded lower semicontinuous function;

(iii) whenever $x_n \to x$, $a_n \to a$, $Q(\cdot | x_n, a_n)$ converges weakly to $Q(\cdot | x, a)$.

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projections X_t, Δ_t on the tth factors S, A describe the state and the action of the tth time of the process ($t \geqq 0$).

A policy is a sequence $\pi = (\pi_0, \pi_1, \cdots)$ such that, for each $t \geqq 0$, π_t is a stochastic kernel on $\mathcal{B}_A \times S \times (A \times S)^t$ with $\pi_t(A(x_t) | x_0, a_0, \cdots, a_{t-1}, x_t) = 1$ for all $(x_0, a_0, \cdots, a_{t-1}, x_t) \in S \times (A \times S)^t$.

Let Π denote the class of policies.

We denote by $T(A|S)$ the set of all stochastic kernels Φ, on $\mathcal{B}_A \times S$ with $\Phi(A(x)|x) = 1$ for all $x \in S$.

A policy $\pi = (\pi_0, \pi_1, \cdots)$ is a randomized stationary policy if there exists a $\Phi \in T(A|S)$ such that $\pi_t(\cdot | x_0, a_0, \cdots, x_t) = \Phi(\cdot | x_t)$ for all $(x_0, a_0, \cdots, x_t) \in S \times (A \times S)^t$ and $t \geqq 0$. Denote the corresponding policy by $\Phi^{(\infty)}$.

For any $D \in \mathcal{B}_S$, we denote by $B(D \to A)$ the set of all Borel measurable functions $u: D \to A$ with $u(x) \in A(x)$ for all $x \in D$.

A randomized stationary policy $\Phi^{(\infty)}$ is called stationary if there exists an $f \in B(S \to A)$ such that $\Phi(\{f(x)\}|x) = 1$ for all $x \in S$. Such a policy will be written by $f^{(\infty)}$.

We denote by Π' and Π'', respectively, the sets of all randomized stationary and stationary policies. For any Borel set X, we denote by $P(X)$ the set of all probability measures on X.

Let $H_t = (X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t)$. It is assumed that, for each $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$, $\text{Prob}(\Delta_t \in D_1 | H_t) = \pi_t(D_1 | H_t)$ and $\text{Prob}(X_{t+1} \in D_2 | H_{t-1}, \Delta_{t-1}, X_t = x, \Delta_t = a) = Q(D_2 | x, a)$ for every $D_1 \in \mathcal{B}_A$ and $D_2 \in \mathcal{B}_S$.

Then, for each $\pi \in \Pi$ and initial state distribution $\nu \in P(S)$, we can define the probability measure P_π^ν on Ω in an obvious way.

We shall consider the following average cost criterion.

For any policy π and initial state distribution $\nu \in P(S)$, let

$$\psi(\nu, \pi) = \limsup_{T \to \infty} E_\pi^\nu \left[\sum_{t=0}^{T-1} c(X_t, \Delta_t) \right] \Big/ T,$$

where E_π^ν is the expectation with respect to P_π^ν.

Let $\psi(\nu) = \inf_{\pi \in \Pi} \psi(\nu, \pi)$ and $\psi^* = \inf_{\nu \in P(S)} \psi(\nu)$. Then we say that $(\nu, \pi) \in P(S) \times \Pi$ is a minimum pair if $\psi(\nu, \pi) = \psi^*$ and $\pi^* \in \Pi$ is optimal if $\psi(x, \pi^*) \leqq \psi(x, \pi)$ for all $x \in S$ and $\pi \in \Pi$, where the initial distribution degenerate at the point x is denoted by x.

2. The existence of minimum pairs in $S \times \Pi''$. In this section we use the hypothesis of Doeblin [7], and give the characterization of minimum pairs.

For any $\Phi \in T(A|S)$, the t-step transition probabilities are defined by

$$Q^{(1)}(\cdot|x, \Phi) = \int Q(\cdot|x, a)\Phi(da|x),$$

$$Q^{(t+1)}(\cdot|x, \Phi) = \int Q^{(t)}(\cdot|x_1, \Phi)Q^{(1)}(dx_1|x, \Phi) \qquad (t \geqq 1).$$

Unless stated otherwise, the following hypothesis holds throughout this paper.

Hypothesis (Doeblin [7]). There is a finite measure γ of sets $D \in \mathcal{B}_S$ with $\gamma(S) > 0$, an integer l and a positive ε, such that

$$Q^{(l)}(D|x, \Phi) \leqq 1 - \varepsilon \quad \text{if } \gamma(D) \leqq \varepsilon, \quad \text{for all } \Phi \in T(A|S) \text{ and } x \in S.$$

For any Borel set X, we denote by $C(X)$ and $C_s(X)$, respectively, the sets of all bounded continuous and lower semicontinuous functions on X. Let $\{s_i\}$ be dense in S and define $f_{ij} \in C(S)$ for $i, j = 1, 2, \cdots$ by

$$f_{ij}(s) = 2(1 - jd(s, s_i)) \vee 0,$$

where d is the metric defined in S and $x \vee y = \max\{x, y\}$. Let $M = \{f_{ij}: i, j = 1, 2, \cdots\}$. Then M is separating, i.e., whenever $P_1, P_2 \in P(S)$ and $\int f_{ij}\, dP_1 = \int f_{ij}\, dP_2$ for $i, j = 1, 2, \cdots$, we have $P_1 = P_2$ (for example, see [8]).

Let us prove the following result using the idea given by Borkar [4], [5].

LEMMA 2.1. *For any* $(\nu, \pi) \in P(S) \times \Pi$, *there exists a* $\mu \in P(R)$ *such that*

$$\int c(x, a)\mu(d(x, a)) \leqq \psi(\nu, \pi), \tag{2.1}$$

$$\int g(x)\mu(d(x, a)) = \int \mu(d(x, a)) \int g(x')Q(dx' \mid x, a) \quad \text{for all } g \in M. \tag{2.2}$$

Proof. For any given $(\nu, \pi) \in P(S) \times \Pi$, we consider $\{X_t, \Delta_t; t = 0, 1, \cdots\}$ governed by P_π^ν. For simplicity put $P = P_\pi^\nu$ and $E = E_\pi^\nu$. Then by the stability theorem of Loève [11], we have, P-almost surely,

$$\lim_{T\to\infty} \left(\sum_{t=0}^{T-1} \{g(X_t) - E[g(X_t)|\mathcal{B}_{t-1}]\} \right) \Big/ T = 0 \quad \text{for all } g \in M, \tag{2.3}$$

where $\mathcal{B}_t = \sigma(X_0, \Delta_0, \cdots, X_t, \Delta_t)$ is the sub-σ-field generated by H_t and Δ_t. Let $\tilde{\psi}_T = \sum_{t=0}^{T-1} c(X_t, \Delta_t)/T$ $(T \geqq 1)$. Then, since $E[\liminf_{T\to\infty} \tilde{\psi}_T] \leqq \psi(\nu, \pi)$, we have

$$P(\liminf_{T\to\infty} \tilde{\psi}_T \leqq \psi(\nu, \pi)) > 0,$$

so that there exists a sample path $\omega \in \Omega$ such that (2.3) and $\liminf_{T\to\infty} \tilde{\psi}_T \leqq \psi(\nu, \pi)$ hold.

Now, let us construct the probability measure for which (2.1) and (2.2) hold by using the fixed sample path $\omega \in \Omega$, which is suppressed for notational convenience. For this $\omega \in \Omega$, there exists a subsequence $\{T_j\}$ for which (2.3) and the following (2.4) hold:

$$\lim_{j\to\infty} \tilde{\psi}_{T_j} \leqq \psi(\nu, \pi). \tag{2.4}$$

For any $D \in \mathcal{B}_R$, define the empirical probability measure μ_T by

$$\mu_T(D) = \sum_{t=0}^{T-1} I_D(X_t, \Delta_t)/T \qquad (T \geqq 0),$$

where, for any set G, I_G is the indicator function of G and $I_G(y)$ is equal to 1 if $y \in G$ and equal to 0 otherwise.

Since R is compact, $P(R)$ is also compact in the topology of weak convergence [3], so that there exists a probability measure $\mu \in P(R)$ such that μ_{T_j} converges weakly to μ along a subsequence (also called $\{T_j\}$ by abuse of notation). We shall show that this μ has the desired property.

By the definition of μ_T, (2.3) and (2.4) are rewritten, respectively, by

$$\lim_{j\to\infty}\left[\int g(x)\mu_{T_j}(d(x,a)) - \int \mu_{T_j-1}(d(x,a)) \int g(x')Q(dx'|x,a)\right] = 0 \tag{2.5}$$

for all $g \in M$

and

$$\lim_{j\to\infty} \int c(x,a)\mu_{T_j}(d(x,a)) \leqq \psi(\nu,\pi). \tag{2.6}$$

By the weak continuity of Q, for each $g \in M$, $\int g(x')Q(dx'|x,a)$ is continuous in $(x,a) \in S \times A$, so that, as $j \to \infty$ in (2.5), we get (2.2). Also, since $c \in C_s(S \times A)$, there exists a nondecreasing sequence $\{\phi_k'\} \subset C(S \times A)$ for which $\phi_k \to c$ as $k \to \infty$.

By (2.6), we have for each $k \geqq 1$,

$$\lim_{j\to\infty} \int \phi_k(x,a)\mu_{T_j}(d(x,a)) = \int \phi_k(x,a)\mu(d(x,a)) \leqq \psi(\nu,\pi).$$

As $k \to \infty$ in the above, using the monotone convergence theorem we get (2.1). □

THEOREM 2.1. *For any* $(\nu,\pi) \in P(S) \times \Pi$, *there exists a* $(\nu_0, \Phi^{(\infty)}) \in P(S) \times \Pi'$ *such that*

$$\psi(\nu_0, \Phi^{(\infty)}) \leqq \psi(\nu,\pi). \tag{2.7}$$

Proof. For $(\nu,\pi) \in P(S) \times \Pi$, let $\mu \in P(R)$ be such that (2.1) and (2.2) hold. Then, we can decompose the probability measure μ into $\nu_0 \in P(S)$ and $\Phi \in T(A|S)$ such that

$$\mu(D_1 \times D_2) = \int_{D_1} \Phi(D_2|x)\nu_0(dx) \text{ for any } D_1 \in \mathcal{B}_S \text{ and } D_2 \in \mathcal{B}_A$$

(for example, see [1]).

Let $Q(\cdot|x,\Phi) = \int Q(\cdot|x,a)\Phi(da|x)$. Then since M is separating, by (2.2) we have

$$\nu_0(\cdot) = \int Q(\cdot|x,\Phi)\nu_0(dx),$$

which means that ν_0 is a stationary absolute probability measure for the Markov process induced by $\{Q(\cdot|x,\Phi)\}$. Hence, we have $\psi(\nu_0,\Phi^{(\infty)}) = \int c(x,a)\mu(d(x,a))$, so by (2.1) we get (2.7). □

Here we can give the main result, which states that there exists a minimum pair in $S \times \Pi''$.

THEOREM 2.2. *There exists* $C \in \mathcal{B}_S$ *with* $\gamma(C) > \varepsilon$ *and* $\bar{f}^{(\infty)} \in \Pi''$ *such that* $(x, \bar{f}^{(\infty)}) \in S \times \Pi''$ *is a minimum pair and* $Q(C|x,\bar{f}(x)) = 1$ *for all* $x \in C$.

Proof. Let $\{\varepsilon_n\}$ be such that $\varepsilon_n > 0$ and $\varepsilon_n \to 0$ as $n \to \infty$. From the definition of ψ^*, there exists a sequence $\{(\nu^n,\pi^n)\} \subset P(S) \times \Pi$ for which it holds that

$$\psi(\nu^n,\pi^n) \leqq \psi^* + \varepsilon_n \quad \text{for all } n \geqq 1. \tag{2.8}$$

From Lemma 2.1, for each (ν^n, π^n) there exists a $\mu^n \in P(R)$ satisfying (2.1) and (2.2) with respect to μ^n.

Also, from (2.1) and (2.8) we have

$$\int c(x, a)\mu^n(d(x, a)) \leqq \psi^* + \varepsilon_n \quad \text{for all } n \geqq 1. \tag{2.9}$$

From the compactness of $P(R)$, without loss of generality there exists a $\mu \in P(R)$ such that μ^n converges weakly to $\mu \in P(R)$ and μ satisfies (2.2). Also, for a nondecreasing sequence $\{\phi_k\} \subset C(S \times A)$ with $\phi_k \to c$ as $k \to \infty$, we have, from (2.9),

$$\int \phi_k(x, a)\mu^n(d(x, a)) \leqq \psi^* + \varepsilon_n.$$

As $n \to \infty$ and $k \to \infty$ in the above, using the monotone convergence theorem we get

$$\int c(x, a)\mu(d(x, a)) \leqq \psi^*. \tag{2.10}$$

Since μ satisfies (2.2), if we decompose μ into $\nu \in P(S)$ and $\Phi \in T(A|S)$, ν is a stationary absolute probability measure for $\{Q(\cdot|x, \Phi)\}$. Thus, by (2.10) we get $\psi(\nu, \Phi^{(\infty)}) \leqq \psi^*$, which implies from the definition that

$$\psi(\nu, \Phi^{(\infty)}) = \psi^*. \tag{2.11}$$

Now we consider the Markov process induced by $\{Q(\cdot|x, \Phi)\}$. Then by the theory of Markov processes, under the hypothesis of Doeblin [7], we can define a decomposition of the state space S into a transient set and a finite number of ergodic sets $C_1, C_2, \cdots, C_k$ with $\gamma(C_j) > \varepsilon$ for all $1 \leqq j \leqq k$.

Here we define a $\gamma_j \in P(C_j)(1 \leqq j \leqq k)$ by

$$\gamma_j(\cdot) = \lim_{T \to \infty} P^{(T)}(\cdot|x, \Phi) \quad \text{for each } x \in C_j,$$

where $P^{(T)}(\cdot|x, \Phi) = \sum_{t=0}^{T-1} Q^{(t)}(\cdot|x, \Phi)/T$ and $Q^{(0)}(\cdot|x, \Phi) = I_{\{x\}}(\cdot)$. Also we have $\nu(\cdot) = \int P^{(T)}(\cdot|x, \Phi)\nu(dx)$ for all $T \geqq 1$.

Hence, as $T \to \infty$ in the above, using the dominated convergence theorem we get

$$\nu(\cdot) = \sum_{j=1}^{k} \gamma_j(\cdot)\nu(C_j). \tag{2.12}$$

Let $c(x, \Phi) = \int c(x, a)\Phi(da|x)$. Then, by definition of $\psi(x, \Phi^{(\infty)})$, we have that, for any $x \in C_j$,

$$\begin{aligned} \psi(x, \Phi^{(\infty)}) &= \limsup_{T \to \infty} \int c(x', \Phi) P^{(T)}(dx'|x, \Phi) \\ &= \int_{C_j} c(x', \Phi)\gamma_j(dx'). \end{aligned} \tag{2.13}$$

On the other hand,

$$\begin{aligned} \psi(\nu, \Phi^{(\infty)}) &= \int c(x, \Phi)\nu(dx) \\ &= \sum_{j=1}^{k} \nu(C_j) \int_{C_j} c(x, \Phi)\gamma_j(dx) \quad \text{from (2.12)} \\ &= \sum_{j=1}^{k} \nu(C_j)\psi(x_j, \Phi^{(\infty)}) \quad \text{for any } x_j \in C_j \qquad (1 \leqq j \leqq k) \quad \text{from (2.13)}, \end{aligned}$$

so that since $\psi(x, \Phi^{(\infty)}) \geqq \psi^*$ for all $x \in S$, it holds from (2.11) that for C_j with $\nu(C_j) > 0$,

$$\psi(x, \Phi^{(\infty)}) = \psi^* \quad \text{for all } x \in C_j. \tag{2.14}$$

Now we prove that $\Phi^{(\infty)}$ above can be replaced by a stationary policy. First we show that for each C_j with $\nu(C_j) > 0$, there exists an $f_j \in B(C_j \to A)$ and a Borel measurable function v_j such that

$$Q(C_j \mid x, f_j(x)) = 1, \tag{2.15}$$

$$v_j(x) + \psi^* \geqq c(x, f_j(x)) + \int_{C_j} v_j(x') Q(dx' \mid x, f_j(x)) \quad \text{for all } x \in C_j. \tag{2.16}$$

The inequality above will be used to show that $f_j^{(\infty)}$ is optimal for all initial states belonging to C_j.

Using the theory of Markov process again, we find that, if ${}_1C_j, {}_2C_j, \cdots, {}_dC_j$ are the cyclically moving classes in C_j and $x \in {}_1C_j$, the following holds:

$$\lim_{t \to \infty} Q^{(td+m-1)}(\cdot \mid x, \Phi) = {}_m\gamma_j(\cdot) \quad \text{and} \quad \gamma_j(\cdot) = \sum_{m=1}^{d} {}_m\gamma_j(\cdot)/d. \tag{2.17}$$

For $T \geqq 1$ and $x \in C_j$, let

$$v_j^T(x) = \sum_{t=0}^{[T/d]d-1} E_{\Phi^{(\infty)}}^{x} (c(X_t, \Delta_t) - \psi^*),$$

where for a real number z, $[z]$ is the largest integer equal to or less than z. Then, from (2.13) and (2.14) we can rewrite, for any $x \in {}_1C_j$,

$$\begin{aligned} v_j^T(x) &= \sum_{t=0}^{[T/d]d-1} \int c(x', \Phi)\{Q^{(t)}(dx' \mid x, \Phi) - \gamma_j(dx')\} \\ &= \sum_{t=0}^{[T/d]-1} \left(\sum_{m=1}^{d} \int c(x', \Phi)\{Q^{(td+m-1)}(dx' \mid x, \Phi) - {}_m\gamma_j(dx')\} \right). \end{aligned}$$

So, since the convergence in (2.17) is uniform and exponentially fast and c is bounded, $\lim_{T \to \infty} v_j^T(x)$ exists and is finite. Let $v_j(x) = \lim_{T \to \infty} v_j^T(x)$ for each $x \in C_j$.

Then by the definition of v_j it holds that

$$v_j(x) + \psi^* = \int \Phi(da \mid x) \left\{ c(x, a) + \int v_j(x') Q(dx' \mid x, a) \right\}. \tag{2.18}$$

Since C_j is an invariant set, $\int Q(C_j \mid x, a)\Phi(da \mid x) = 1$ for all $x \in C_j$, so that for each $x \in C_j$

$$Q(C_j \mid x, a) = 1, \qquad \Phi(\cdot \mid x) \text{ a.s.} \tag{2.19}$$

Here, let

$$\bar{R} = \left\{ (x, a) \,\middle|\, Q(C_j \mid x, a) = 1, x \in C_j, a \in A(x), v_j(x) + \psi^* \geqq c(x, a) + \int v_j(x') Q(dx' \mid x, a) \right\}.$$

Then, by (2.18) and (2.19), $\Phi(\bar{R}_x \mid x) > 0$ for all $x \in C_j$, where

$$\bar{R}_x = \{a \in A(x) \mid (x, a) \in \bar{R}\}.$$

Thus, from the selection theorem of [2], there exists an $f_j \in B(C_j \to A)$ such that $f_j(x) \in \bar{R}_x$ for all $x \in C_j$.

Clearly (2.15) and (2.16) hold for this f_j.

Let $C = \bigcup C_j$, where the union is over all $j \in \{i \mid \nu(C_i) > 0\}$. For any $f \in B(S \to A)$, if we define $\bar{f} \in B(S \to A)$ by

$$\bar{f}(x) = \begin{cases} f_j(x) & \text{if } x \in C_j \text{ and } \nu(C_j) > 0, \\ f(x) & \text{otherwise,} \end{cases}$$

it holds from (2.15) and (2.16) that

$$Q(C \mid x, \bar{f}(x)) = 1 \quad \text{for all } x \in C, \tag{2.20}$$

$$v(x) + \psi^* \geqq c(x, \bar{f}(x)) + \int v(x') Q(dx' \mid x, \bar{f}(x)) \quad \text{for all } x \in C, \tag{2.21}$$

where $v(x) = v_j(x)$ if $x \in C_j$ with $\nu(C_j) > 0$.

Let us show that $\psi(x, \bar{f}^{(\infty)}) \leqq \psi^*$ for all $x \in C$ by the same way as the proof of Theorem 7.6 in [13].

For $x \in C$, from (2.21) it holds that, for each $t \geqq 0$,

$$E^x_{\bar{f}^{(\infty)}}[c(X_t, \Delta_t)] \leqq \psi^* + E^x_{\bar{f}^{(\infty)}}[v(X_t) - v(X_{t+1})].$$

Therefore,

$$\begin{aligned} \psi(x, \bar{f}^{(\infty)}) &= \limsup_{T \to \infty} \sum_{t=0}^{T-1} E^x_{\bar{f}^{(\infty)}}[c[X_t, \Delta_t)]/T \\ &\leqq \psi^* + \limsup_{T \to \infty} (v(x) - E^x_{\bar{f}^{(\infty)}}[v(X_t)]/T) \\ &= \psi^*. \end{aligned}$$

□

Theorem 2.2 does not establish the existence of a minimum pair for all initial states—only for those in a Borel subset C of S. At the same time, it shows the existence of an optimal stationary policy for these restricted initial states. In the next section, the extension of these results to all of S is done assuming additional conditions.

3. Optimal stationary policies. In this section we discuss the existence of optimal stationary policies under the following two conditions.

Condition A (Reachability). For any $x \in S$ and $D \in \mathcal{B}_S$ with $\gamma(D) > \varepsilon$, there exists a $\pi \in \Pi$ such that

$$P^x_\pi\left(\bigcup_{t=0}^{\infty} \{X_t \in D\}\right) = 1,$$

where ε and γ are as in the hypothesis of Doeblin.

Condition B. One of the following two conditions is satisfied:

(B1) For γ as in the hypothesis of Doeblin, $\gamma(\partial D) = 0$ if $\gamma(D) > 0$, where ∂D is the boundary of D.

(B2) For each $D \in \mathcal{B}_S$ with $\gamma(D) > \varepsilon$, $Q(D \mid x, a)$ is continuous in $(x, a) \in S \times A$.

Now we can state the following theorem.

THEOREM 3.1. *Under Conditions* A *and* B, *there exists an optimal stationary policy.*

To prove Theorem 3.1 we need several lemmas.

Let C and $\bar{f} \in B(S \to A)$ be given in Theorem 2.2 and $F = S - C$.

Let $\tilde{t} = \inf\{t \geqq 0 \mid X_t \in C\}$, where $\inf \phi = \infty$.

For any $x \in F$, let π be such that $P^x_\pi(\bigcup_{t=0}^{\infty} \{X_t \in C\}) = 1$, whose existence is guaranteed by Condition A.

For this π, we define $\pi^* = (\pi_0^*, \pi_1^*, \cdots)$ by, for each $D \in \mathcal{B}_A$,

$$\pi_t^*(D \mid H_t) = \begin{cases} I_D(\bar{f}(X_t)) & \text{if } t \geqq \tilde{t}, \\ \pi(D \mid H_t) & \text{if } t < \tilde{t}. \end{cases}$$

We note that $P_{\pi^*}^x(\bigcup_{t=0}^{\infty} \{X_t \in C\}) = 1$.

LEMMA 3.1. $\psi(x, \pi^*) = \psi^*$.

Proof. For simplicity set $E = E_{\pi^*}^x$. Since $\psi(x, \bar{f}^{(\infty)}) = \psi^*$ for all $x \in C$, it follows from the definition of π^* that for any $k \geqq 0$ with $P_{\pi^*}^x(\tilde{t} = k) > 0$,

$$\begin{aligned} &\limsup_{T\to\infty} E\left[\sum_{t=k}^{T-1} c(X_t, \Delta_t) \,\middle|\, \tilde{t} = k\right] \Big/ T \\ &= \lim_{T\to\infty} E\left[\sum_{t=k}^{T-1} c(X_t, \Delta_t) \,\middle|\, \tilde{t} = k, X_k \in C\right] \Big/ (T-k) \qquad (3.1) \\ &= \psi^*. \end{aligned}$$

Also,

$$\begin{aligned} \psi(x, \pi^*) \leqq &\sum_{k=0}^{\infty} E\left[\limsup_{T\to\infty} E\left[\sum_{t=0}^{(T-1)\wedge k} c(X_t, \Delta_t) \,\middle|\, \tilde{t} = k\right] \Big/ T\right] \\ &+ \sum_{k=0}^{\infty} E\left[\limsup_{T\to\infty} E\left[\sum_{t=(T-1)\wedge k+1}^{T-1} c(X_t, \Delta_t) \,\middle|\, \tilde{t} = k\right] \Big/ T\right], \end{aligned}$$

where $a \wedge b = \min\{a, b\}$.

Since c is bounded, the first term of the right-hand side of the above inequality is zero, so that by (3.1) we get

$$\psi(x, \pi^*) \leqq \psi^* P_{\pi^*}^x\left(\bigcup_{t=0}^{\infty} \{X_t \in C\}\right) = \psi^*. \qquad \square$$

For $u \in C_s(F)$, we define Ku, for each $x \in F$, by

$$Ku(x) = \inf_{a \in A(x)} \left\{\psi^* Q(C^* \mid x, a) + \int_F u(x') Q(dx' \mid x, a)\right\}, \qquad (3.2)$$

where $C^* = \mathring{C}$ under (B1) and $C^* = C$ under (B2), and $\mathring{C}$ is the set of interior points of C.

LEMMA 3.2. *$u \in C_s(F)$ implies $Ku \in C_s(F)$.*

Proof. Let $u \in C_s(F)$. Under (B1), $\gamma(\mathring{C}) \geqq \varepsilon$, and from the weak continuity of Q, $\liminf_{x' \to x, a' \to a} Q(\mathring{C} \mid x', a') \geqq Q(\mathring{C} \mid x, a)$, which means that $Q(\mathring{C} \mid x, a) \in C_s(F \times A)$.

Also, $\int_F u(x') Q(dx' \mid x, a) \in C_s(F \times A)$, so that $Ku \in C_s(F)$ (cf. [1], [12]).

Under (B2), similarly $Ku \in C_s(F)$. $\square$

LEMMA 3.3. *Let v be a bounded Borel measurable function on F such that*

$$v = Kv. \qquad (3.3)$$

Then, $v(x) \leqq \psi^*$ for all $x \in F$.

Proof. For $x \in F$ and $\pi^* \in \Pi$ as in Lemma 3.1, let

$$B_T(x, \pi^*) = \psi^* P_{\pi^*}^x(X_t \in C^* \text{ for some } t \leqq T).$$

For simplicity set $P = P^x_{\pi^*}$ and $E = E^x_{\pi^*}$. We have

$$B_T(x, \pi^*) = E\left[\sum_{t=0}^{T-1} \psi^* Q(C^* | X_t, \Delta_t) I_{\{X_0 \in F, \cdots, X_t \in F\}}\right]$$

$$\geqq E\left[\sum_{t=0}^{T-1} (v(X_t) - E[v(X_{t+1}) I_{\{X_{t+1} \in F\}} | X_t, \Delta_t]) \cdot I_{\{X_0 \in F, \cdots, X_t \in F\}}\right] \quad \text{from (3.3)}$$

$$\geqq v(x) - H \cdot P(X_t \in F \text{ for all } 0 \leqq t \leqq T) \quad \text{for some } H > 0.$$

Thus, as $T \to \infty$ in the formula above, we get

$$\lim_{T\to\infty} B_T(x, \pi^*) = \psi^* P\left(\bigcup_{t=0}^{\infty} \{X_t \in C\}\right) = \psi^* \geqq v(x). \qquad \square$$

Proof of Theorem 3.1. For any $\alpha > \psi^*$, let $v_0 \equiv \alpha$ and $v_{n+1} = Kv_n (n \geqq 0)$. Then we observe from Lemma 3.2 that $v_n \in C_s(F)$ and $v_n \geqq v_{n+1} \geqq \psi^*$ for all $n \geqq 0$. Let $v = \lim_{n\to\infty} v_n$. Then $Kv \leqq Kv_n = v_{n+1} (n \geqq 1)$, which implies $Kv \leqq v$. On the other hand, for any $\eta > 0$,

$$\eta + Kv > \psi^* Q(C | x, a) + \int v(x') Q(x' | x, a) \quad \text{for some } a \in A(x). \tag{3.4}$$

By the monotone convergence theorem,

$$\lim_{n\to\infty} \int v_n(x') Q(dx' | x, a) = \int v(x') Q(dx' | x, a).$$

Hence, from (3.4) we have

$$\eta + Kv > \psi^* Q(C | x, a) + \int v_n(x') Q(dx' | x, a)$$

$$\geqq v_{n+1}(x) \quad \text{for all } n \geqq N \text{ and some } N.$$

As $n \to \infty$ in the above, $\eta + Kv(x) \geqq v(x)$, which gets $Kv(x) \geqq v(x)$ when $\eta \to 0$.

From the discussion above, we have $v = Kv$.

By Lemma 3.2, we get $v \leqq \psi^*$. Hence, if we let $\tilde{k} = \min\{k | v_k(x) < \alpha\}$ and $F_k = \{x \in F | \tilde{k}(x) = k\}$ for each $k \geqq 1$, it holds that

$$F = \bigcup_{k=1}^{\infty} F_k. \tag{3.5}$$

By the selection theorem (cf. [1], [12]), for each $k \geqq 1$ there exists an $f_k \in B(F_k \to A)$ with

$$v_k(x) = \psi^* Q(C | x, f_k(x)) + \int v_{k-1}(x') Q(dx' | x, f_k(x)) \quad \text{for all } x \in F_k. \tag{3.6}$$

Now, considering (3.5), we define $\bar{\bar{f}} \in B(S \to A)$ by

$$\bar{\bar{f}}(x) = \begin{cases} \bar{f}(x) & \text{if } x \in C, \\ f_k(x) & \text{if } x \in F_k \quad \text{for some } k \geqq 1. \end{cases}$$

Here, we consider the stationary Markov process induced by $\{Q(\cdot | x, \bar{\bar{f}}(x))\}$. For each $k \geqq 1$, it holds from the definition of F_k that $Q(F_{k-1} | x, \bar{\bar{f}}(x)) > 0$ for all $x \in F_k$, so that by (3.5) F is a transient set, which implies $P^x_{\bar{\bar{f}}^{(\infty)}}(\bigcup_{t=0}^{\infty} \{X_t \in C\}) = 1$ for all $x \in F$, where $F_0 = C$.

Using Lemma 3.1, we get $\psi(x, \bar{\bar{f}}^{(\infty)}) = \psi^*$ for all $x \in S$. $\square$

Remark. We introduce the following condition to consider the unichain case.

Condition C. For any $f \in B(S \to A)$, the Markov process induced by $\{Q(\cdot | x, f(x))\}$ has only one ergodic set.

Examination of the discussion in §§ 2 and 3 shows that there exists an optimal stationary policy under Conditions B and C. These are extensions of the results obtained in [14], [15] by approximating the average cost problem by the discounted cost problems.

4. Further results. In this section we derive the general results under the following continuity condition.

Condition D. The following hold:

(D1) For any $G \in \mathcal{B}_S$, $Q(G|\cdot) \in C(R)$.

(D2) For any sequence $\{x_n\}$ and x in S with $x_n \to x$ as $n \to \infty$ and any $a \in A(x)$, there exists a sequence $\{a_n\}$ with $a_n \in A(x_n)$ and $a_n \to a$ as $n \to \infty$.

THEOREM 4.1. *Suppose that Condition* D *holds; then there exists a decomposition of* S:

$$S = F \cup S_1 \cup \cdots \cup S_r, \quad F \in \mathcal{B}_S, \quad S_i \in \mathcal{B}_S, \tag{4.1}$$
$$S_i \cap S_j = \phi \quad (i \neq j), \quad F \cap S_j = \phi,$$

and a stationary policy $f^{(\infty)}$ *and constants* $\alpha_1, \alpha_2, \cdots, \alpha_r$ *with the following properties*:

(i) $\psi(x, f^{(\infty)}) = \psi(x)$ *for all* $x \in S^*$ *and* $\psi(x) = \alpha_i$ *for each* $x \in S_i$ $(1 \leqq i \leqq r)$, *where* $S^* = \bigcup_{i=1}^{r} S_i$.

(ii) *Each* $S_i (1 \leqq i \leqq r)$ *is invariant for the Markov process induced by* $f^{(\infty)}$.

(iii) $\psi(x, f^{(\infty)}) \leqq \psi(x, \pi)$ *for all* $x \in F$ *and any* $\pi \in \Pi$ *with*

$$P_\pi^x(X_t \in S^* \text{ for some } t \geqq 0) = 1. \tag{4.2}$$

Proof. By Theorem 2.2, there exists a constant α_1, a set $S_1 \in \mathcal{B}_S$ with $\gamma(S_1) > \varepsilon$, and a stationary policy $\bar{f}_1^{(\infty)}$ for which $\psi(x, \bar{f}_1^{(\infty)}) = \psi(x) = \alpha_1$ and $Q(S_1|x, \bar{f}_1(x)) = 1$ for all $x \in S_1$, where γ and ε are as in the hypothesis of Doeblin. We note that S_1 corresponds to C in Theorem 2.2. Put $D_0 = S_1$ and $D = S - D_0$. If we set, for each $j \geqq 1$, $D_j = \{x \in D - \bigcup_{i=1}^{j-1} D_i \mid Q(D_{j-1}|x, a) > 0 \text{ for some } a \in A(x)\}$, D_j is open. In fact, let $\{x_n\}$ be any sequence such that $x_n \notin D_j$ and $x_n \to x$ as $n \to \infty$. Then for any $a \in A(x)$, by (D2) there exists a sequence $\{a_n\}$ with $a_n \in A(x_n)$ and $a_n \to a$ as $n \to \infty$. From (D1), we have $Q(D_{j-1}|x, a) = \lim_{n\to\infty} Q(D_{j-1}|x_n, a_n) = 0$.

In the case where $D = \bigcup_{j=1}^{\infty} D_j$, from the selection theorem (for example, [1]), there exists an $\hat{f} \in B(D \to A)$ such that for each $j \geqq 1$

$$Q(D_{j-1}|x, \hat{f}(x)) = \max_{a \in A(x)} Q(D_{j-1}|x, a) \quad \text{for all } x \in D_j.$$

Here we define f by

$$f(x) = \begin{cases} \bar{f}_1(x) & \text{if } x \in S_1, \\ \hat{f}(x) & \text{if } x \in D. \end{cases}$$

Clearly, $P_{f^{(\infty)}}^x(X_t \in S_1 \text{ for some } t \geqq 0) = 1$ for all $x \in D$, so that by Lemma 3.1, $\psi(x, f^{(\infty)}) = \alpha_1$. Thus, putting $F = D$, the proof is complete.

In the case where $D \neq \bigcup_{j=1}^{\infty} D_j$, let $G = D - \bigcup_{j=1}^{\infty} D_j$. Then, using (D2) we can prove that $Q(\bar{G}|x, a) = 1$ for all $x \in \bar{G}$ and $a \in A(x)$, where $\bar{G}$ is the closure of G. We note from the hypothesis of Doeblin that $\gamma(\bar{G}) > \varepsilon$.

Here, we apply Theorem 2.2 to the sub-Markov decision process with the restricted state space $\bar{G}$. Then there exists a constant α_2, a set $S_2 \in \mathcal{B}_{\bar{G}}$ with $\gamma(S_2) > \varepsilon$ and a stationary policy $\bar{f}_2^{(\infty)}$ for which $\psi(x, \bar{f}_2^{(\infty)}) = \psi(x) = \alpha_2$ and $Q(S_2|x, \bar{f}_2(x)) = 1$ for all $x \in S_2$.

For any given $g \in B(S \to A)$, we define $\bar{f} \in B(S \to A)$ by

$$\bar{f}(x) = \begin{cases} \bar{f}_i(x) & \text{if } x \in S_i, \quad i = 1, 2, \\ g(x) & \text{otherwise.} \end{cases}$$

Then we observe that $\psi(x, \bar{f}^{(\infty)}) = \alpha_i$ if $x \in S_i$, $i = 1, 2$. Now, putting $D_0 = S_1 \cup S_2$ and $D = S - D_0$, we repeat the discussion above. Since $\gamma(S) < \infty$, by repeating this method successively we come to the conclusion that there exists a decomposition (4.1) and $\hat{f} \in B(S \to A)$ satisfying (i) and (ii). If we let $G_0 = S^*$ and for each $j \geqq 1$,

$$G_j = \left\{ x \in F - \bigcup_{i=1}^{j-1} G_i \mid Q(G_{j-1} \mid x, a) > 0 \text{ for some } a \in A(x) \right\},$$

we have

$$F = \bigcup_{j=1}^{\infty} G_j, \tag{4.3}$$

where $F = S - S^*$.

For $u \in C(\bar{F})$, we define Uu by, for each $x \in \bar{F}$,

$$Uu(x) = \inf_{a \in A(x)} U(x, a, u), \tag{4.4}$$

where

$$U(x, a, u) = \sum_{i=1}^{r} \alpha_i Q(S_i \mid x, a) + \int_F u(x') Q(dx' \mid x, a) \quad \text{for } x \in \bar{F} \text{ and } a \in A(x). \tag{4.5}$$

Let $v_0 \equiv \max\{\alpha_i; i = 1, \cdots, r\}$ and $v_{n+1} = Uv_n (n \geqq 0)$. Then, clearly $v_n \geqq v_{n+1} (n \geqq 1)$ and $v_n \in C(\bar{F})$. If we put $v = \lim_{n \to \infty} v_n$, $v \in C(\bar{F})$ from Dini's Theorem.

Also, similarly as the proof of Theorem 3.1, we get

$$v(x) = Uv(x) \quad \text{for all } x \in F. \tag{4.6}$$

Now we define, for each $x \in F$ and $\pi \in \Pi$,

$$B(x, \pi) = \lim_{T \to \infty} B_T(x, \pi),$$

where

$$B_T(x, \pi) = \sum_{i=1}^{r} \alpha_i P_\pi^x \quad (X_t \in S_i \text{ for some } 0 \leqq t \leqq T).$$

Let π be any policy satisfying (4.2). Then it is easily proved that

$$\psi(x, \pi) = B(x, \pi) \geqq v(x) \quad \text{for all } x \in F. \tag{4.7}$$

Here, again by the selection theorem, there exists an $f^* \in B(F \to A)$ such that $v(x) = U(x, f^*(x), v)$ for all $x \in F$. Using this f^*, we define $f \in B(S \to A)$ by

$$f(x) = \begin{cases} \hat{f}(x) & \text{if } x \in S^*, \\ f^*(x) & \text{if } x \in F. \end{cases}$$

From (4.3) we observe that $f^{(\infty)}$ satisfies (4.2). Therefore,

$$v(x) = \lim_{T \to \infty} U^T v(x) \geqq \lim_{T \to \infty} B_T(x, f^{(\infty)}) = B(x, f^{(\infty)}) = \psi(x, f^{(\infty)}),$$

which implies (iii) from (4.7). □

Acknowledgments. The author expresses his thanks to the referees and an associate editor for their many valuable comments and suggestions that improved the presentation of the material.

REFERENCES

[1] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control—The Discrete Time Case*, Academic Press, New York, 1978.

[2] D. BLACKWELL AND C. RYLL-NARDZEWSKI, *Non-existence of everywhere proper conditional distributions*, Ann. Math. Statist., 34 (1963), pp. 223-225.

[3] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[4] V. S. BORKAR, *Controlled Markov chains and stochastic networks*, SIAM J. Control Optim., 21 (1983), pp. 652-666.

[5] ———, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965-978.

[6] H. DEPPE, *On the existence of average optimal policies in semi-regenerative decision models*, Math. Oper. Res., 9 (1984), pp. 558-575.

[7] J. L DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[8] S. N. ETHIER AND T. G. KURTZ, *Markov Processes, Characterization and Convergence*, John Wiley, New York, 1986.

[9] R. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.

[10] M. KURANO, *Markov decision processes with a Borel measurable cost function—the average case*, Math. Oper. Res., 11 (1986), pp. 309-320.

[11] M. LOÈVE, *Probability Theory*, Second edition, Van Nostrand, Princeton, NJ, 1960.

[12] A. MAITRA, *Discounted dynamic programming on compact metric space*, Sankhyā Ser. A, 30 (1968), pp. 211-216.

[13] S. M. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

[14] ———, *Arbitrary state Markovian decision processes*, Ann. Math. Statist., 39 (1968), pp. 2118-2122.

[15] H. C. TIJMS, *On dynamic programming with arbitrary state space, compact action space and the average return as criterion*, Report BW 55/75, Math. Centrum, Amsterdam, 1975.

[16] P. L. SCHWEITZER, *On the existence of relative value for undiscounted multichain Markov decision processes*, J. Math. Anal. Appl., 102 (1984), pp. 449-455.

[17] W. H. M. ZIJM, *The optimality equation in multichain denumerable state Markov decision processes with the average cost criterion: the bounded cost case*, Statist. Decisions, 3 (1985), pp. 143-165.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 308-329, March 1989

005

REGULAR SYNTHESIS FOR THE LINEAR TIME-OPTIMAL CONTROL PROBLEM WITH CONVEX CONTROL CONSTRAINTS*

MARGARÉTA HALICKÁ†

Abstract. The problem of existence of a regular synthesis for the linear time-optimal control problem with convex control constraints is studied. A regular synthesis on the whole reachable set cannot be established for this problem by direct use of Brunovsky's general existence theorem. This is in accord with the example of a nonsubanalytic reachable set due to Lojasiewicz and Sussmann [S. Lojasiewicz, Jr. and H. J. Sussmann, "Some examples of reachable sets and optimal cost functions that fail to be subanalytic," *SIAM J. Control Optim.*, 23 (1985), pp. 584-598]. A closed subset H of the reachable set K that has Lebesgue measure zero is constructed and the existence of a regular synthesis on $K - H$ is proved.

Key words. control theory, linear time-optimal problem, regular synthesis, reachable set, subanalytic set

AMS(MOS) subject classifications. 49C20, 34H05

1. Introduction. The class of linear time-optimal control problems is one of the simplest and most developed classes of optimal control problems. Linear time-optimal control problems having polyhedral control restraint sets and satisfying a normality condition are the first larger class of problems for which the existence of a regular synthesis has been proved [2]. This result was obtained by using the method of subanalytic sets, which has found wide application to other problems in control theory as well.

Using the abstract theorem on the existence of regular synthesis [3] (ERS theorem) the existence of a regular synthesis for the linear-quadratic problem (LQ problem) has been proved in [4]. The control set is a polyhedron and a normality condition is assumed. A certain generalization of the LQ problem is given in [6], which contains a proof of the existence of the regular synthesis for a linear-convex problem (LC problem), the system is linear, the Lagrangian is convex and strictly convex in the control variable, and the convex compact control set is given by analytic inequalities.

All the classes of problems [2], [4], [6] mentioned have one common property derived from the application of the method of subanalytic sets: they admit a subanalytic regular synthesis, e.g., all the cells of the regular synthesis are subanalytic sets. From this property it follows that the whole reachable set is subanalytic.

The results achieved in control theory using the method of subanalytic sets seem to indicate that subanalyticity is a property of other, larger classes of control problems. Hence the paper by Lojasiewicz and Sussmann [8] was a surprise. It presents some examples of rather simple problems that do not possess the property of subanalyticity. This is first of all an example of a linear control system with a convex control set. In this example the reachable set is not subanalytic, and therefore the optimal control problem for this system does not admit a subanalytic regular synthesis on the whole reachable set. So, in this case, the applicability of the method of subanalytic sets in control theory reaches its limits.

Here we deal with the same class of linear time-optimal control problems with a convex, compact control set given by analytic inequalities. For this class of problems we investigate the existence of a subanalytic regular synthesis. From [8], however, it

* Received by the editors September 8, 1986; accepted for publication (in revised form) May 13, 1988. This paper is part of the author's doctoral thesis written at Comenius University.

† Institute of Applied Mathematics and Computing, Comenius University, Mlynská dolina, Bratislava, Czechoslovakia.

is evident that a subanalytic regular synthesis does not exist on the whole reachable set. Therefore, using the ERS theorem, we construct a closed subset H of the reachable set K, and we prove the existence of a regular synthesis on $K-H$, the cells being subanalytic subsets of $K-H$ (§§ 1-6). In §§ 7 and 8 we prove that the Lebesgue measure of H is zero, i.e., $K-H$ almost exhausts K.

2. The linear time-optimal control problem. We consider the time-optimal control problem given by the linear system

$$\dot{x} = Ax + Bu \tag{1}$$

where A, B are $n \times n$ and $n \times m$ matrices, respectively, $n \geqq m$, rank $B > 1$, and the $n \times nm$ matrix $(B, AB, A^2B, \cdots, A^{n-1}B)$ has rank n.

Here the class of admissible controllers consists of measurable functions defined on intervals $[0, T]$, $T > 0$, with values in the set $U \subset R^m$. The initial point is $x_0 \in R^n$ and the target point is $\hat{x} = 0 \in R^n$.

The control restraint set U is a compact, convex set, $0 \in \operatorname{int} U$ and it is assumed to be of the form

$$U = \{u \in R^m / g^i(u) \leqq 0,\ i \in S,\ c^{j*}u \leqq d^j,\ j \in P\}$$

(* stands for transposition), $S = \{1, \cdots, s\}$, $P = \{1, \cdots, p\}$, where

(a) $g^i : R^m \to R$, $i \in S$, are analytic functions, $c^j \in R^m$, $d^j \in R^1$;

(b) $M_i(u) = (\partial^2 g^i(u)/\partial u^2) > 0$ for $u \in U$, $i \in S$;

(c) If for $\hat{u} \in U$ we have $g^i(\hat{u}) = 0$ for $i \in P_1 \subseteq P$ and $c^{j*}\hat{u} = d^j$ for $j \in S_1 \subseteq S$, then the vectors $(\partial g^i(u)/\partial u)^*$, c^j, $i \in P_1, j \in S_1$, are linearly independent;

(d) Among the inequalities defining U there are no redundant ones, i.e., for every $i \in S$ there exists a $u \in R^m$ such that $g^i(u) > 0$, $g^k(u) \leqq 0$, $c^{j*}u \leqq d^j$ for $k \in S - \{i\}, j \in P$, and for every $j \in P$ there exists a $u \in R^m$ such that $c^{j*}u > d^j$, $c^{k*}u \leqq d^k$, $g^i(u) \leqq 0$ for $k \in P - \{j\}$, $i \in S$.

Before presenting the last assumption about the set U we introduce some notation. We define

$$U_{IJ} = \{u \in R^m / g^i(u) = 0,\ i \in I,\ g^i(u) < 0,\ i \notin I,\ c^{j*}u = d^j,\ j \in J,\ c^{j*}u < d^j,\ j \notin J\}$$

for every index set $I \times J$, where $I \subseteq S$, $J \subseteq P$, and $|I| + |J| \leqq m$, $|I|$ being the cardinality of I.

The sets U_{IJ} can be empty for some I, J. For $I = \varnothing$, $J = \varnothing$, U_{IJ} is the interior of U. We shall call the index set $I \times J$ admissible if $|I| + |J| \leqq m$ and U_{IJ} is not empty and denoted by IJ. Instead of U_{IJ} we shall sometimes write U_I, U_J, or U_0, respectively, when $J = \varnothing$, $I = \varnothing$, or $IJ = \varnothing$.

Now we return to the formulation of the assumption for U.

(e) For every U_J, J admissible, there are $z^1, z^2 \in \bar{U}_J$ such that the vectors $B(z^1 - z^2), AB(z^1 - z^2), \cdots, A^{n-1}B(z^1 - z^2)$ are linearly independent.

For given $x_0 \in R^n$ we will denote by $\mathrm{LT}(x_0)$ the time-optimal control problem just formulated.

Let us note that for our purpose it is sufficient if the assumptions (a), (b) are valid on some neighbourhood of U.

Assumption (e) will also be called the normality condition. From this condition it follows that the LT problem is normal in the sense of [9].

A controller $u(t)$ and its response $x(t)$, $t \in [0, T]$, will be called extremal for the LT problem (with respect to a nonzero solution $\psi(t)$ of the adjoint equation $\dot{\psi} = -A^*\psi$)

if the triple x, u, ψ satisfies the maximum condition

$$\psi(t)^* Bu(t) = \max_{u \in U} \psi(t)^* Bu$$

almost everywhere on $[0, T]$.

We recall some well-known properties of the linear time-optimal control problem that are valid under our assumptions (cf. [7], [9]).

(LT1) If $0 \leqq T_1 < T_2$, then $K(T_1) \subseteq \operatorname{int} K(T_2)$ ($K(T)$ is the set of all initial points x from which (1) can be steered to the origin using an admissible controller defined on the interval $[0, T]$).

(LT2) The reachable set K (i.e., the set of all points x from which (1) can be steered to the origin using an admissible controller defined on some finite interval) is an open subset of R^n.

(LT3) For given $x \in K$ there exists a unique extremal controller $u_x(t)$ defined on $[0, T_x]$. This extremal controller is optimal.

(LT4) The function $T(x)$ associating with every x the minimal time T_x in which (1) can be steered from x to zero is continuous on K.

Our aim is to investigate the existence of a subanalytic regular synthesis for the problem LT(x), $x \in K$. As in the LQ problem and the LC problem we use the ERS theorem.

3. The solution of the maximum condition. To verify the hypothesis of the ERS theorem we must express the solution of the maximum condition as a function of ψ. Since U is compact, for each $\psi \in R^n - \{0\}$ there exists a $w_\psi \in U$ such that the linear function $H_\psi(u) = \psi^* Bu$ has its maximum in w_ψ, i.e.,

$$\psi^* B w_\psi = \max_{u \in U} \psi^* Bu. \tag{2}$$

Contrary to the LC problem, the solution w_ψ of condition (2) is not uniquely determined by the value ψ in general. For example, if $\psi \in \operatorname{Ker} B^*$ or $B^*\psi$ is from the convex cone generated by the vectors $c^j, j \in J$, we can choose for w_ψ, $u \in U$, or $u \in U_J$, respectively.

We denote by $\bar{w}(\psi)$ the set of all solutions w_ψ of the condition (2) for given ψ.

In the sequel we decompose the space of the adjoint variables into two sets W_A and W_B. The solution of (2) will be uniquely determined by ψ on W_A, which enables us to use the procedure from [6] on W_A. First, we introduce some notation and prove some lemmas.

If a property V_i (V_j, respectively) holds for every $i \in I$ ($j \in J$), then we shall write $V_I(V_J)$ instead of $V_i, i \in I$ ($V_j, j \in J$).

In the same way as in the LC problem [6] we can prove the following lemma about an analytic stratification of U.

LEMMA 1. *The family of the sets U_{IJ}, IJ admissible, is an analytic stratification of U.*

COROLLARY 1. *Let IJ be an admissible set. Then* $\dim U_{IJ} = m - (|I| + |J|)$ *and U_{IJ} has at most a finite number of components.*

The proof follows from the linear independence of the vectors $(\partial g^i(u)/\partial u)^*$, c^j, $i \in I, j \in J$. The finiteness of the numbers of components follows from the semianalyticity of U_{IJ}.

Further, we define

$$W_{IJ} = \{\psi \in R^n - \{0\} / \bar{w}(\psi) \cap U_{IJ} \neq \varnothing\} \tag{3}$$

for every IJ admissible. It is easy to see that for $IJ = \emptyset$ the set $U_{IJ} = \operatorname{int} U$ and $W_{IJ} = \operatorname{Ker} B^* - \{0\}$. Instead of W_{IJ} we shall sometimes write W_I, W_J, or W_0, respectively, when $J = \emptyset$, $I = \emptyset$, or $IJ = \emptyset$. However, any statement about IJ, I, or J admissible will be understood to hold for the case $IJ = \emptyset$, $I = \emptyset$, or $J = \emptyset$ as well.

Because of the positive homogeneity of H in ψ the sets W_{IJ}, IJ admissible, completed by the point zero, are cones. These sets cover $R^n - \{0\}$. In contrast to the LC problem the sets W_{IJ}, IJ admissible, do not partition the space $R^n - \{0\}$.

Example. Let the restraint set U be given by the inequalities

$$u_1 \leqq \tfrac{1}{2}, \qquad u_1^2 + u_2^2 + u_3^2 \leqq 1,$$

and let $n = m = 3$ and B be an identity matrix. We denote

$$U_{12} = \{u \in R^3 / u_1 = \tfrac{1}{2},\ u_1^2 + u_2^2 + u_3^2 = 1\},$$

$$U_1 = \{u \in R^3 / u_1 < \tfrac{1}{2},\ u_1^2 + u_2^2 + u_3^2 = 1\},$$

$$U_2 = \{u \in R^3 / u_1 = \tfrac{1}{2},\ u_1^2 + u_2^2 + u_3^2 < 1\}.$$

The sets U_{12}, U_1, U_2 form an analytic stratification of U defined in the general case above. The corresponding set W_{12} is the closed circular cone with the vertex in the origin of the space of adjoint variables and with the axis ψ_1. The set W_2 is the positive half-axis ψ_1, W_1 is the open circular cone $R^3 - W_{12} - \{0\}$.

As with the LC problem, we can prove a lemma about subanalyticity of W_{IJ} and the necessary and sufficient condition for $\psi \in W_{IJ}$ in our LT problem. The proofs of these lemmas are the same as the corresponding ones for the LC problem.

LEMMA 2. *For every IJ admissible the set W_{IJ} is a subanalytic subset of R^n.*

LEMMA 3. *Let $\psi \in R^n - \{0\}$, IJ admissible. Then $\psi \in W_{IJ}$ if and only if there exist a $u \in U_{IJ}$ and $a^i \geqq 0$, $i \in I$, $b^j \geqq 0$, $j \in J$, such that*

$$-B^*\psi + \sum_{i \in I} a^i \left(\frac{\partial g^i(u)}{\partial u} \right)^* + \sum_{j \in J} b^j c^j = 0. \tag{4}$$

u is a solution of (2) for given ψ.

Remark 1. Let $u \in U_{IJ}$, IJ admissible. By the symbol C_u we denote the convex cone generated by the vectors $(\partial g^i(u)/\partial u)^*$, c^j, $i \in I, j \in J$. Then Lemma 3 can be written in the form

$$B^* W_{IJ} = \bigcup_{u \in U_{IJ}} C_u. \tag{5}$$

COROLLARY 2. *For every J admissible the set W_J is a closed subset of $R^n - \{0\}$.*

Proof. Evidently $B^* W_J = C_u$, where u is an arbitrary chosen element from U_J. The set C_u is closed for every $u \in U_J$ and therefore its B^*-preimage is closed. □

COROLLARY 3. (a) $\dim B^* W_J = |J|$ *for every J admissible.*

(b) *If $|J| = m - k$, then* $\dim W_J \leqq n - k$.

(c) $\dim W_0 < n - 1$.

Proof. Assertion (a) is a corollary of the linear independence of the vectors c^j, $j \in J$, J admissible.

Condition (b) follows from the fact that a linear mapping does not increase the dimension of an analytic manifold.

If $J = \emptyset$, then $|J| = m - m$, and because of $m > 1$ we have $\dim W_0 = \dim W_J \leqq n - m < n - 1$. □

COROLLARY 4. *For every IJ, J' admissible, such that $J' \subseteq J$ the inclusion $W_{J'} \subseteq W_{IJ}$ holds.*

Proof. Let $\psi \in W_{J'}$; then according to Lemma 3 there exist $b^j \geqq 0$, $j \in J'$ such that

$$-B^*\psi + \sum_{j \in J} b^j c^j = 0.$$

We take $a^i = 0$, $i \in I$, $b^j = 0$, $j \in J - J'$, and an arbitrary $u \in U_{IJ}$. Then, for $\psi, a^i, i \in I, b^j, j \in J, u$, condition (4) holds and therefore $\psi \in W_{IJ}$. This completes the proof. □

For every IJ admissible we define the set V_{IJ} by

$$V_{IJ} = \begin{cases} W_{IJ} - W_J & \text{if } |I| \neq 0, \\ W_{IJ} - \bigcup_{J' \subset J} W_{J'} & \text{if } |I| = 0. \end{cases} \tag{6}$$

As in the case of W's instead of V_{IJ} we shall sometimes write V_I, V_J, or V_0, respectively, when $J = \emptyset$, $I = \emptyset$, or $IJ = \emptyset$.

LEMMA 4. *For every $\psi \in V_{IJ}$, IJ admissible, IJ such that either $|I| \neq 0$ or $|J| = m$, there is a unique solution w_ψ of the maximum condition* (2).

Proof. Let $\psi \in V_{IJ}$, IJ admissible, $|I| \neq 0$. Then there exists a $w_\psi \in U_{IJ}$ such that

$$\psi^* B w_\psi = \max_{u \in U} \psi^* B u. \tag{7}$$

Let $\hat{w}_\psi \in U_{I'J'}$, $\hat{w}_\psi \neq w_\psi$, be such that

$$\psi^* B \hat{w}_\psi = \max_{u \in U} \psi^* B u. \tag{8}$$

Then it follows immediately from (7), (8) and from the convexity of U that for every $\lambda \in [0, 1]$

$$\psi^* B(\lambda w_\psi + (1-\lambda)\hat{w}_\psi) = \max_{u \in U} \psi^* B u. \tag{9}$$

Hence $\tilde{w}_\psi(\lambda) = \lambda w_\psi + (1-\lambda)\hat{w}_\psi$ is a solution of the maximum condition for every $\lambda \in [0, 1]$, i.e., $\tilde{w}_\psi(\lambda)$ is a maximum of a linear function on a compact set. Since $\psi \in V_{IJ}$, $\psi \notin W_0$ and therefore $w_\psi(\lambda)$ is from the boundary of U for every $\lambda \in [0, 1]$. From the definition of U (the strict convexity and the linearity of the inequalities defining U) it can be seen that if the segment $\tilde{w}_\psi(\lambda)$, $\lambda \in [0, 1]$, is from the boundary of U, then there is a J'' admissible such that $\tilde{w}_\psi(\lambda) \in U_{J''}$ for every $\lambda \in (0, 1)$ and $J'' \subseteq J'$, $J'' \subseteq J$. Consequently, $\psi \in W_{J''} \subseteq W_J$, which contradicts the definition of V_{IJ}.

Let $\psi \in V_J$, $|J| = m$; then there exists a unique $w_\psi \in U_J$ since U_J is a vertex of U. Now, because of the same argument as in the case of $|I| \neq 0$ we obtain the uniqueness of $w_\psi \in U$. □

COROLLARY 5. *The sets V_{IJ}, IJ admissible, form a partition of $R^n - \{0\}$.*

Proof. Due to the definition of the sets V_{IJ} they cover $R^n - \{0\}$. First, we prove that if IJ is admissible such that either $|I| \neq 0$ or $|J| = m$, then $V_{IJ} \cap V_{I'J'} = \emptyset$ for every $I'J'$ admissible, $IJ \neq I'J'$.

Let $\psi \in V_{IJ} \cap V_{I'J'}$. Then there are $w_\psi^1, w_\psi^2 \in \bar{w}(\psi)$ such that $w_\psi^1 \in U_{IJ}$, $w_\psi^2 \in U_{I'J'}$. According to Lemma 1 we have $U_{IJ} \cap U_{I'J'} = \emptyset$ and therefore $w_\psi^1 \neq w_\psi^2$, which contradicts Lemma 4.

Now we prove that if J is admissible, $J < m$, then $V_J \cap V_{J'} = \emptyset$ for every J' admissible, $J' \neq J$, $|J'| < m$. Let $\psi \in V_J \cap V_{J'}$. Then there exist $w_\psi^1 \in U_J$, $w_\psi^2 \in U_{J'}$ and $w_\psi^1 \neq w_\psi^2$. Due to the definitions of U and w_ψ there is J'' admissible, $J'' \subseteq J'$, $J'' \subseteq J$ such that $\psi \in W_{J''}$. This contradicts the definition of V_J.

We denote

$$W_A = \bigcup V_{IJ}, \tag{10}$$

where the union is taken over all IJ admissible such that either $|I| \neq 0$ or $|J| = m$. Further, we denote

$$W_Y = (R^n - \{0\}) - W_A. \tag{11}$$

Since the sets V_{IJ}, IJ admissible, form a partition of $R^n - \{0\}$ the following relation holds:

$$W_Y = \bigcup_{|J|<m} V_J = \bigcup_{|J|<m} W_J. \tag{12}$$ □

COROLLARY 6. *The set W_A is an open subset of $R^n - \{0\}$.*

This follows immediately from (10)-(12) and Corollary 2.

Now it is easy to see the geometric meaning of the definitions of W_A and W_Y: W_A is the complement of W_Y in $R^n - \{0\}$; W_Y is the union of the polyhedral cones W_J of dimensions smaller than n; W_J is the B^*-pre-image of the convex cone generated by the vectors $c^j, j \in J$.

From Lemma 4 and the definition of W_A the uniqueness of solutions of the maximum condition for $\psi \in W_A$ follows. This enables the definition of a function $w(\psi)$ on the set W_A by the formula $w(\psi) \in \bar{w}(\psi)$.

LEMMA 5. *The function $w(\psi)$ is continuous on its domain of definition.*

The proof follows much the same lines as that of the similar lemma in [6] and therefore is omitted.

Remark 2. The uniqueness of the solution of the maximum condition (2), Lemma 3, and the linear independence of the vectors $(\partial g^I(u)/\partial u)^*$, c^J result in the uniqueness of the numbers a^I, b^J from Lemma 3. For this reason we can define the functions $a^I(\psi)$, $b^J(\psi)$ on V_{IJ} for every IJ admissible.

4. Properties of the function $w(\psi)$. In the previous lemmas we have shown that the function $w(\psi)$ possesses properties on the set W_A similar to the analogous function $w(x, \psi)$ mentioned in the LC problem. The fact that it is continuous and satisfies the maximum condition has enabled us to prove in the LC problem that the functions $w_I(x, \psi)$ can be extended as analytic functions to neighbourhoods of the closures of the sets W_I. In this section we prove a similar property for the function $w(\psi)$ on V_{IJ}. The closures and neighbourhoods of the sets V_{IJ}, IJ admissible, where either $|I| \neq 0$ or $|J| = m$, will be considered as subsets of W_A. We denote by cl_A the relative closure in W_A.

The proofs of the next three lemmas follow the same pattern as the proofs of the similar lemmas in [6]. Therefore we do not present their proofs.

LEMMA 6. *Let IJ be admissible such that $|I| \neq 0$. Then* int $V_{IJ} \neq \emptyset$.

LEMMA 7. *Let IJ be admissible such that $I \neq \emptyset$. Then*

$$\mathrm{cl}_A V_{IJ} \subseteq \bigcup_{I'J' \supseteq IJ} V_{I'J'} \tag{13}$$

where the union is taken over all $I'J'$ admissible such that $I'J' \supseteq IJ$.

LEMMA 8. *Let IJ be admissible, $|I| \neq 0$, let $\psi \in \mathrm{cl}_A V_{IJ}$. Then there is an admissible set $I'J'$, $I'J' \supseteq IJ$ such that $\psi \in V_{I'J'}$ and the functions $a^{I'}$, $b^{J'}$ defined for $I'J'$ satisfy*

$$a^i(\psi) = 0 \quad \text{for } i \in I' - I,$$

$$b^j(\psi) = 0 \quad \text{for } j \in J' - J.$$

Remark 3. Because of Lemma 8 we extend the functions a^I, b^J defined on V_{IJ} to $\mathrm{cl}_A V_{IJ}$ by the formulas

$$a^i(\psi) = a'^i(\psi), \qquad i \in I,$$
$$b^j(\psi) = b'^j(\psi), \qquad j \in J$$

for $\psi \in \mathrm{cl}_A V_{IJ} - V_{IJ}$. Here a'^i, b'^j are, respectively, the ith component of the function a'^I from Lemma 8 and the jth component of b'^J from Lemma 8.

LEMMA 9. *Let IJ be admissible,* $|I| \neq 0$. *Suppose for* $\hat{\psi} \in R^n$, $\hat{u} \in U$, $\hat{a}^I \geqq 0$, $\hat{b}^J \geqq 0$ *the conditions*

$$-B^*\hat{\psi} + \sum_{i \in I} \hat{a}^i \left(\frac{\partial g^i(\hat{u})}{\partial u}\right)^* + \sum_{j \in J} \hat{b}^j c^j = 0, \tag{14}$$

$$g^I(\hat{u}) = 0, \tag{15}$$

$$c^{J*}\hat{u} = d^J \tag{16}$$

hold, where $\hat{a}^i > 0$ *for at least one* $i \in I$. *Then there exist a neighbourhood* O *of* $\hat{\psi}$ *and analytic functions* $\alpha^I(\psi), \beta^J(\psi), u(\psi)$ *defined on* O *such that* $\alpha^I(\hat{\psi}) = \hat{a}^I$, $\beta^J(\hat{\psi}) = \hat{b}^J$, $u(\hat{\psi}) = \hat{u}$ *and the equations*

$$-B^*\psi + \sum_{i \in I} \alpha^i(\psi) \left(\frac{\partial g^i(u(\psi))}{\partial u}\right)^* + \sum_{j \in J} \beta^j(\psi) c^j = 0,$$
$$g^I(u(\psi)) = 0, \qquad c^{J*}u(\psi) = d^J$$

hold for every $\psi \in O$ *and* $\alpha^i(\psi) > 0$.

Proof. Let us define the function

$$F : R^n \times R^{|I|} \times R^{|J|} \times R^m \to R^m \times R^{|I|} \times R^{|J|}$$

by the formula

$$F(\psi, a^I, b^J, u) = \begin{pmatrix} -B^*\psi + \sum_{i \in I} a^i \left(\frac{\partial g^i(u)}{\partial u}\right)^* + \sum_{j \in J} b^j c^j \\ g^I(u) \\ c^{J*}u - d^J \end{pmatrix}$$

Then $F(\hat{\psi}, \hat{a}^I, \hat{b}^J, \hat{u}) = 0$. We denote

$$M_I(\hat{u}) = \sum_{i \in I} \hat{a}^i M_i(\hat{u}), \qquad G_{IJ}(\hat{u}) = \left(\left(\frac{\partial g^I(\hat{u})}{\partial u}\right)^*, c^J\right),$$

where $M_i(u)$ has been defined in § 2. The matrices $M_I(\hat{u})$ and $G_{IJ}(\hat{u})$ are of the types $n \times m$ and $m \times (|I| + |J|)$, respectively. The matrix $M_I(\hat{u})$ is positive definite and according to [5]

$$\det \frac{\partial F(\hat{\psi}, \hat{a}^I, \hat{b}^J, \hat{u})}{\partial(a^I, b^J, u)} = \det \begin{pmatrix} M_I(\hat{u}) & G_{IJ}(\hat{u}) \\ G_{IJ}(\hat{u})^* & 0 \end{pmatrix}$$
$$= \det M_I(\hat{u}) \cdot \det \left(G_{IJ}(\hat{u})^* M_I(\hat{u})^{-1} G_{IJ}(\hat{u})\right) \neq 0.$$

The statement of our theorem follows immediately from the implicit function theorem. □

Remark 4. From the previous lemma we obtain

$$\frac{\partial(u, a^I, b^J)}{\partial \psi}(\hat{\psi}) = \begin{pmatrix} M_I(\hat{u}) & G_{IJ}(\hat{u}) \\ G_{IJ}(\hat{u})^* & 0 \end{pmatrix}^{-1} \begin{pmatrix} B^* \\ 0 \end{pmatrix}. \tag{17}$$

Then, using Frobenius' formula [5] we obtain

$$\frac{\partial u}{\partial \psi} = (M_I^{-1} - M_I^{-1} G_{IJ} (G_{IJ}^* M_I^{-1} G_{IJ})^{-1} G_{IJ}^* M_I^{-1}) B^*, \tag{18}$$

$$\frac{\partial (a^I, b^J)}{\partial \psi} = (G_{IJ}^* M_I^{-1} G_{IJ})^{-1} G_{IJ}^* M_I^{-1} B^*. \tag{19}$$

THEOREM 1. *For given V_{IJ}, IJ admissible, where either $|I| \neq 0$ or $|J| = m$, there exists a neighbourhood B_{IJ} of $\mathrm{cl}_A V_{IJ}$ in W_A and an analytic function $w_{IJ}(\psi)$ defined on B_{IJ} such that $w_{IJ}(\psi) = w(\psi)$ for every $\psi \in \mathrm{cl}_A V_{IJ}$.*

Proof. The statement of this theorem for IJ admissible, $|I| \neq 0$, can be proved from Lemmas 6–9 by a procedure similar to the one by which Theorem 3 of [6] is proved from Lemmas 5–8 in the LC problem.

If $|J| = m$, then the function $w(\psi)$ is constant on V_J and therefore the assertion of this theorem is trivial. □

Now, the example from § 3 will show that for V_{IJ}, $|J| < m-1$, the function $w_{IJ}(\psi)$ defined on V_{IJ} cannot always be continuously extended to the neighbourhood of the closure of V_{IJ} in $R^n - \{0\}$.

Example. Let U_1, U_2, U_{12}, W_1, W_2, W_{12}, B, m, n be as in the example of § 3. Then, by (2), $w_\psi = u = (u_1, u_2, u_3)$ can be expressed as a function of $\psi = (\psi_1, \psi_2, \psi_3)$ on $V_{12} = W_{12} - W_2$, i.e.,

$$u_1 = \tfrac{1}{2}, \qquad u_1^2 + u_2^2 + u_3^2 = 1,$$

$$\psi_1 = 2au_1 + b, \quad \psi_2 = 2au_2, \quad \psi_3 = 2au_3.$$

Solving this system, we obtain

$$u_1 = \tfrac{1}{2},$$

$$u_2 = \psi_2 \sqrt{3} / (2\sqrt{\psi_2^2 + \psi_3^2}),$$

$$u_3 = \psi_3 \sqrt{3} / (2\sqrt{\psi_2^2 + \psi_3^2}).$$

It is easy to see that, for example,

$$\lim_{\substack{\psi_2 \to 0 \\ \psi_3 \to 0}} u_2(\psi)$$

does not exist and therefore the function $w_{12}(\psi) = (u_1(\psi), u_2(\psi), u_3(\psi))$ cannot be continuously extended to W_{12}.

The example shows that the LT problem does not in general satisfy Assumption 2 of the ERS theorem and therefore the existence of a regular synthesis cannot be proved directly using this theorem on the whole reachable set K. In § 6 we shall define a subset H of K such that the existence of a regular synthesis can be proved on $K - H$.

First, we shall extend the set W_A to a set $\tilde{W}_A$ which will include the sets V_J, $|J| = m-1$ and we shall prove an analogue to Theorem 1 on $\tilde{W}_A$. We denote

$$\tilde{W}_A = W_A \cup \bigcup_{|J| = m-1} V_J, \tag{20}$$

$$W_Z = \bigcup_{|J| < m-1} V_J = \bigcup_{|J| < m-1} W_J. \tag{21}$$

Because of Corollary 5,

$$\tilde{W}_A = R^n - \{0\} - W_Z \tag{22}$$

and from Corollary 3 it follows that

$$\dim W_Z < n-1. \tag{23}$$

LEMMA 10. *The sets* $\mathrm{cl}_{\tilde{A}} V_{IJ}$, *IJ admissible such that either* $|I| \neq 0$ *or* $|J| = m$, *cover* $\tilde{W}_A$.

Proof. Because of Corollary 5 the sets V_{IJ}, *IJ* admissible and such that either $|I| \neq 0$ or $|J| = m$, form a partition of W_A. Let J be admissible, $|J| = m-1$. Then U_J is an open segment of a line; because of the boundedness of U and Lemma 1 there exists $I'J'$ admissible such that $J \subseteq I'J'$ and $|I'| + |J'| = m$. It follows from Corollary 4 and the definition of the sets V_{IJ} that $V_J \subseteq \mathrm{cl}_A V_{I'J'}$, which completes the proof. □

COROLLARY 7. *For every component X of* V_{IJ}, *IJ admissible, where either* $|I| \neq 0$ or $|J| = m$, *there exist a neighbourhood* O_{IJ} *of* $\mathrm{cl}_{\tilde{A}} X$ *in* $\tilde{W}_A$ *and an analytic function* w_{IJ} *defined on* O_{IJ} *such that* $w_{IJ}(\psi) = w(\psi)$, $\psi \in X$.

Proof. Let *IJ* be such that $|I| + |J| = m$. Then $\dim U_{IJ} = 0$ and $w(\psi)$ is constant on each component of V_{IJ}. Therefore, $w_{IJ}(\psi) = w(\psi)/V_{IJ}$ can be extended as constant to a neighbourhood of the relative closure of V_{IJ} in $\tilde{W}_A$.

Let V_{IJ} be such that $|I| \neq 0$ and $|J| < m-1$. Due to the definitions of W_A and $\tilde{W}_A$, $\mathrm{cl}_A V_{IJ} = \mathrm{cl}_{\tilde{A}} V_{IJ}$ and the statement is a direct corollary of Theorem 1. □

5. Finiteness of the number of switchings. In this section we demonstrate a connection between $w(\psi)$ and extremal controllers and we prove that these controllers have a locally finite number of switchings.

The function $w(\psi)$ has not been defined on W_Y (for W_Y see (12)). Because of the normality condition and analyticity of $e^{-A^*t}\psi$, for every $\psi \in R^n - \{0\}$ and for a finite interval $\mathcal{T}$ there is at most a finite number of points $t_i \in \mathcal{T}$ such that $e^{-A^*t_i}\psi \in W_Y$. This enables us to define the function $v(\psi, t): R^n \times R \to R^m$ be the formula

$$v(\psi, t) = w(e^{-A^*t}\psi).$$

Moreover, if we take into account the continuity of $w(\psi)$ on the complement W_A of W_Y (Lemma 5), we obtain the following two properties of $v(\psi, t)$.

LEMMA 11. *The function* $v(\psi, t)$ *is a measurable function of t for every given* ψ.

LEMMA 12. *For every* $\psi_0 \in R^n - \{0\}$ *there exists a set* $H \subset R$ *of measure zero such that for each* $t \in R - H$, $v(\cdot, t)$ *is continuous at* ψ_0.

We define the function $F: (R^n - \{0\}) \times [0, \infty) \to R$ by the formula

$$F(\psi, T) = -\int_0^T e^{-At} B v(\psi, t)\, dt. \tag{24}$$

THEOREM 2. *The function* $F(\psi, T)$ *is continuous. The range of F is the reachable set K.*

Proof. The continuity of F is a consequence of Lemmas 11 and 12 and the boundedness of $v(\psi, t)$. Since F is continuous in T uniformly with respect to ψ, the function $F(\psi, T)$ is continuous.

The inclusion range $F \subseteq K$ being trivial, we prove $K \supseteq$ range F. Let $x \in K$. Due to (LT3) there exists an extremal controller $u_x(t)$ steering x to the origin in time T. Let $\psi(t)$ be a corresponding adjoint solution. Since $v(\psi(0), t)$ is uniquely defined as a solution of the maximum condition for every $t \in [0, T]$ outside a finite set, $v(\psi(0), t) = u_x(t)$ holds almost everywhere on $[0, T]$. As $u_x(t)$ steers x to the origin on $[0, T]$, we

have

$$x = -\int_0^T e^{-At} B u_x(t)\, dt = -\int_0^T e^{-At} B v(\psi(0), t)\, dt.$$

Now it is easy to see that $(\psi(0), T) \in F^{-1}(x)$ and that the range of F contains K. □

From the theorem just proved and from the uniqueness of the extremal controllers (LT3), we obtain the following corollary.

COROLLARY 8. *If $(\psi, T) \in F^{-1}(x)$, then $v(\psi, \cdot)$ as a function of $t \in [0, T]$ for fixed ψ is the unique extremal controller steering x to the origin.*

THEOREM 3. *For every compact subset M of the reachable set K there exists a $\nu(M) > 0$ such that an extremal controller of the* LT(x) *problem has at most ν switching points for every $x \in M$.*

Proof. Due to Corollary 8 it suffices to prove that if $x \in M$, $(\psi, T) \in F^{-1}(x)$, then the function $v(\psi, \cdot)$, $t \in [0, T]$, has at most $\nu(M)$ switching points, where $\nu(M)$ is independent of x. That means we need to verify only that for fixed ψ the curve $e^{-A^* t}\psi$ crosses at most $\nu(M)$ times from one set V_{IJ} to another.

Let M be a compact subset of K. Its pre-image under the continuous map F is a closed subset of the space $(R^n - \{0\}) \times R$ of the variables (ψ, T). Due to (LT4) the set $F^{-1}(M)$ is bounded by a constant T_1. We denote

$$M_1 = \{\psi \in R^n / (\psi, T) \in F^{-1}(M) \text{ for some } T, |\psi| = 1\},$$

$$M_2 = \{e^{-A^* t}\psi \in R^n / \psi \in M_1, 0 \leqq t \leqq T_1\}.$$

The sets M_1 and M_2 are compact.

Consider the partition $\mathcal{P}$ of the space $R^n - \{0\}$ of ψ into the sets V_{IJ}, IJ admissible. The vector field $-A^*\psi$ and the partition $\mathcal{P}$ satisfy the assumption of Theorem II in [11]. By this theorem there exists $\nu(M_2) > 0$ such that every trajectory $e^{-A^* t}\psi$, $\psi \in M_1$, $t \in [0, T_1]$ crosses from one member of the partition $\mathcal{P}$ to another at most $\nu(M)$ times. This completes the proof. □

6. Domain of the existence of a regular synthesis. In § 4 we showed that in general the LT problem does not satisfy Assumption 2 of the ERS theorem on the whole space $R^n - \{0\}$ of the adjoint variables. We found, however, that on the open set $\tilde{W}_A \subset R^n - \{0\}$ the assumption is satisfied. In this section we construct an open subset of the reachable set K, such that the extremal responses going from the points of this set to the origin stay in it, and such that the corresponding solutions $\psi(t)$ of the adjoint equation are from $\tilde{W}_A$. We shall prove the existence of a regular synthesis on this subset.

We have

$$C = \{\psi \in R^n / |\psi| = 1 \text{ and there exists a } t \in R \text{ such that } e^{-A^* t}\psi \in W_Z\}. \tag{25}$$

Remark 5. From the maximum condition (2) it is easy to see that if w is a solution of (2) for given ψ, then w is a solution of (2) for every $c\psi$, $c > 0$. Therefore $w(\psi) = w(c\psi)$, $c > 0$. Since the solutions $\psi(t) = e^{-A^* t}\psi$ are homogeneous in ψ, then also $v(c\psi, t) = w(e^{-A^* t} c\psi) = w(c\, e^{-A^* t}\psi) = w(e^{-A^* t}\psi) = v(\psi, t)$. Therefore $F(c\psi, T) = F(\psi, T)$ as well, where F is the function defined by (24).

Since the set W_Z completed by the origin is a cone, we have the following:

(a) $F(cC \times [0, T)) = F(C \times [0, T))$, where $cC = \{c\psi / \psi \in C\}$, $c > 0$;

(b) If $x \in K - F(C \times [0, \infty))$ such that $F(\psi, T) = x$, then there does not exist a $t \in [0, \infty)$ such that $e^{-A^* t}\psi \in W_Z$.

We denote

$$H = F(C \times [0, \infty)). \tag{26}$$

LEMMA 13. *The set H is a closed subset of the reachable set K for the* LT *problem.*

Proof. Let $\{x_n\} \to x$, $x_n \in H$, $x \in K$. Let $\{\psi_n, T_n\}$ be a sequence of points from $C \times [0, \infty)$ such that $F(\psi_n, T_n) = x_n$. From Corollary 8 and (LT3) it follows that $T_n = T(x_n)$, i.e., T_n is a minimal time for LT(x_n). Because of (LT4) there exists $T_0 = T(x)$ such that $\{T_n\} \to T_0$. Since C is compact there exists a subsequence $\{\psi_{n_k}\}$ of $\{\psi_n\}$ convergent in C. Let $\{\psi_{n_k}\} \to \psi_0$, $\psi_0 \in C$. Due to the continuity of F it follows that $x = \lim_{n\to\infty} x_n = \lim_{k\to\infty} F(\psi_{n_k}, T_{n_k}) = F(\psi_0, T_0)$. Therefore $x \in H$. □

As a direct consequence of Lemma 13 and (LT2) we obtain Corollary 9.

COROLLARY 9. *The set $K - H$ is an open subset of R^n.*

Since $K - H$ is an open subset of R^n and the extremal trajectories of the points from $K - H$ stay in $K - H$, we can consider on it the existence of a regular synthesis in the sense of the definition of [3].

In this section we shall prove the existence of a regular synthesis for the LT(x), $x \in K - H$. We use the property that Assumption 2 of the ERS theorem can be replaced by an assumption that the functions $w_{IJ}(\psi)$ can be extended as analytic functions along the solutions of the adjoint system in our case (here the target point is from the interior of K).

THEOREM 4. *The* LT *problem admits a regular synthesis in the sense of* [3] *on $K - H$.*

Proof. The theorem will be proved using the ERS theorem.

Assumption 1 of the ERS theorem is satisfied trivially.

For the sets N_i and the functions w_i of Assumption 2 we take the sets $N_{IJ} = (K - H) \times \bar{V}_{IJ}$, IJ admissible such that either $|I| \neq 0$ or $|J| = m$, and the functions $w'_{IJ}(x, \psi) = w_{IJ}(\psi)$ (for w_{IJ} see Corollary 7). The sets N_{IJ} and the functions w'_{IJ} satisfy all the required assumptions except the one stating that they can be extended as analytic functions to a neighbourhood of N_{IJ} in $R^n \times (R^n - \{0\})$. This hypothesis from [3] is, however, unnecessary. The proof of the ERS theorem [3] uses only the fact that w_i can be extended as an analytic function to a neighbourhood of N_i along the flow of the differential equation [3, Eq. (12)]. This means in our case that the function $w_{IJ}(\psi)$, defined and analytic on V_{IJ}, can be extended as an analytic function to the set $C_{IJ} = \{\psi \in R^n - \{0\} / \exists \psi_0 \in \bar{V}_{IJ}, t > 0$, such that $\psi = e^{-A^* t}\psi_0\}$. This holds by Corollary 7.

The first part of Assumption 3 of the ERS theorem is satisfied due to (LT3). The second can be proved in the same way as the similar assertion in [6].

Assumptions 4–6 of the ERS theorem follow from the argument of Theorem 3 and (LT4). Assumption 7 can be proved in the same way as in [6].

In the following two sections we shall prove that the set H is "small." For the criterion of "smallness" we shall take its Lebesgue measure to be zero. The proof consists of two steps. The first step proves that the function $w(\psi)$ can be extended as a locally Lipschitz continuous function in a weakened sense to a part of W_Z. The second step uses a countable covering of $C \times [0, \infty)$ such that the function $F(\psi, T)$ is locally Lipschitz continuous on every element of this covering.

7. Lipschitz continuity of $w(\psi)$. In this section we study the Lipschitz continuity properties of $w(\psi)$ on the set C. Some properties of $w(\psi)$ on $C - W_Z$ follow from the previous sections: $w(\psi)$ is continuous on $C - W_Z$ (Lemma 5) and $w(\psi)$ is analytic on every $V_{IJ} \cap C$ and can be analytically extended to some neighbourhood of the closure of $V_{IJ} \cap C$ in $C - W_Z$ (Theorem 1). The following lemma is a direct consequence of [11, Lemma 1] and the mentioned properties of $w(\psi)$ on $C - W_Z$.

LEMMA 14. *The function $w(\psi)$ is locally Lipschitz continuous on $C - W_Z$, i.e., for every compact subset K of $R^n - W_Z$ there is an $L > 0$ such that $|w(\psi_1) - w(\psi_2)| \leqq L|\psi_1 - \psi_2|$ for every $\psi_1, \psi_2 \in K \cap C$.*

In a series of lemmas we shall find an upper bound for $\partial w(\psi)/\partial\psi$ given by (17) and (18).

Let A be the $p \times r$ matrix with elements a_{ij}. We denote by $|A|$ and $\|A\|$, respectively, the Euclidean and operator norms of A, i.e., $|A| = \sqrt{\sum_{i,j} (a_{ij})^2}$ and $\|A\| = \sup_{x \neq 0} (|Ax|/|x|)$. Further, we recall that $M_I(u, a^I) = \sum_{i \in I} a^i M_i(u)$ is an $m \times m$ positive definite matrix for each $a^I \geqq 0$, $a^I \neq 0$, $u \in U$, and $G_{IJ}(u) = ((\partial g^I(u)/\partial u), c^J)$ is an $m \times (|I| + |J|)$ matrix of rank $|I| + |J|$ for every $u \in \bar{U}_{IJ}$.

In what follows we shall denote by P_J the $|J| \times m$ matrix of the orthogonal projection of R^m into the linear hull of $c^j, j \in J$, and by S_J the $(m - |J|) \times m$ matrix of the orthogonal projection of R^m into the orthogonal complement of the linear hull of $c^j, j \in J$. We denote

$$R_J = \begin{pmatrix} P_J \\ S_J \end{pmatrix},$$

i.e., R_J is a regular $m \times m$ matrix.

LEMMA 15. *Let IJ be admissible, $|I| \neq 0$. Let $a^I(\psi)$ be the function defined on $\mathrm{cl}_A V_{IJ}$ in Remark 3. Then there exist positive numbers k_1, k_2 such that the relation*

$$k_1 |S_J B^* \psi| \leqq \sum_{i \in I} a^i(\psi) \leqq k_2 |S_J B^* \psi| \tag{27}$$

holds for each $\psi \in \mathrm{cl}_A V_{IJ}$.

Proof. By Lemmas 3 and 8, for each $\psi \in \mathrm{cl}_A V_{IJ}$ there exist a $u \in \bar{U}_{IJ}$ and an $a^I(\psi), b^J(\psi) \geqq 0$ such that

$$B^* \psi = \sum_{i \in I} a^i \frac{\partial g^i(u)}{\partial u} + \sum_{j \in J} b^j c^j. \tag{28}$$

Since $|I| \neq 0$, $\psi \notin W_0$, and $\psi \notin W_J$, and therefore at least one $a^i > 0$. Because of the linear independence of $(\partial g^I(u)/\partial u), c^J$ for each $u \in \bar{U}_{IJ}$, the rank of the $m \times (|I| + |J|)$ matrix $(\partial g^I(u)/\partial u, c^J)$ is $|I| + |J|$. Due to the regularity of R_J the rank of the matrix

$$\begin{pmatrix} P_J \\ S_J \end{pmatrix} \left(\frac{\partial g^I(u)}{\partial u}, c^J \right) = \begin{pmatrix} P_J \dfrac{\partial g^I(u)}{\partial u} & P_J c^J \\ S_J \dfrac{\partial g^I(u)}{\partial u} & 0 \end{pmatrix}$$

is $|I| + |J|$ and so the rank of the $(m - |J|) \times |I|$ matrix $S_J(\partial g^I(u)/\partial u)$ is $|I|$ for every $u \in \bar{U}_{IJ}$. Let us multiply (28) by the matrix R_J from the left. Then (28) is equivalent to

$$P_J B^* \psi = \sum_{i \in I} a^i P_J \frac{\partial g^i(u)}{\partial u} + \sum_{j \in J} b^j P_J c^j, \tag{29}$$

$$S_J B^* \psi = \sum_{i \in I} a^i S_J \frac{\partial g^i(u)}{\partial u}. \tag{30}$$

Consider the function

$$h(a^I, u) = \sum_{i \in I} a^i S_J \frac{\partial g^i(u)}{\partial u} \tag{31}$$

defined for every $a^I \in R^{|I|}$, $u \in \bar{U}_{IJ}$. Since for $u \in \bar{U}_I$ the rank of $S_J(\partial g^I(u)/\partial u)$ is $|I|$, $h(a^I, u) = 0$ if and only if $a^I = 0$. We divide both sides of (30) by $\sum_{l \in I} a^l$. Then

$$\frac{|S_J B^* \psi|}{\sum_{l \in I} a^l} = \sum_{i \in I} a'^i S_J \frac{\partial g^i(u)}{\partial u} \tag{32}$$

where

$$a'^i = a^i \Big/ \sum_{l \in I} a^l. \tag{33}$$

From (33) it follows that $\sum_{i \in I} a'^i = 1$. The right-hand side of (32) is the value of h in $u \in \bar{U}_{IJ}$ and a'^I such that $\sum_{i \in I} a'^i = 1$. Since h is continuous and the set of all a'^i such that $\sum_{i \in I} a'^i = 1$ and the set $\bar{U}_{IJ}$ are compact, the function h attains its maximum and minimum values on this set. Hence, there exist positive numbers h_1, h_2 such that

$$h_2 \leqq \left| \sum_{i \in I} a'^i S_J \frac{\partial g^i(u)}{\partial u} \right| \leqq h_1 \tag{34}$$

for each $u \in \bar{U}_{IJ}$ and all numbers a'^i such that $\sum_{i \in I} a'^i = 1$. Relation (27) follows from (32)–(34); there, $k_1 = 1/h_1$ and $k_2 = 1/h_2$. □

LEMMA 16. *For each IJ admissible,* $|I| \neq 0$, *there exist constants* $k_1 > 0$, $k_2 > 0$ *such that*

$$|M_I^{-1}(u, a^I)| \leqq k_1 \Big/ \left(\sum_{i \in I} a^i \right), \tag{35}$$

$$|(G_{IJ}^*(u) M_I^{-1}(u, a^I) G_{IJ}(u))^{-1}| \leqq k_2 \sum_{i \in I} a^i \tag{36}$$

for each $u \in \bar{U}_{IJ}$ *and* $a^I \geqq 0$, $a^I \neq 0$.

Proof. Let $\lambda_0(u, a^I)$ and $\lambda_1(u, a^I)$ be, respectively, the smallest and largest eigenvalues of the matrix $M_I(u, a^I)$. Then

$$\lambda_0(u, a^I) \leqq \|M_I(u, a^I)\| = \lambda_1(u, a^I), \tag{37}$$

$$\frac{1}{\lambda_1(u, a^I)} \leqq \|M_I^{-1}(u, a^I)\| = \frac{1}{\lambda_0(u, a^I)}, \tag{38}$$

$$\frac{1}{\lambda_1(u, a^I)} = \min_{x \neq 0} \frac{x^* M_I^{-1}(u, a^I) x}{|x|^2}. \tag{39}$$

We denote by $m_0^i(u)$ and $m_1^i(u)$, respectively, the smallest and the largest eigenvalues of $M_i(u)$ for each $i \in I$. Since $m_0^i(u)$ and $m_1^i(u)$ are continuous functions of u defined on $\bar{U}_{IJ}$ with positive values, there exist

$$m_0^i = \min_{u \in \bar{U}_{IJ}} m_0^i(u) \quad \text{and} \quad m_1^i = \max_{u \in \bar{U}_{IJ}} m_1^i(u).$$

Let us denote $m_0 = \min_{i \in I} m_0^i$ and $m_1 = \max_{i \in I} m_1^i$. Then

$$x^* M_I(u, a^I) x = \sum_{i \in I} x^* a^i M_i(u) x \leqq m_1 |x|^2 \sum_{i \in I} a^i,$$

$$x^* M_I(u, a^I) x \geqq m_0 |x|^2 \sum_{i \in I} a^i$$

from which it follows that

$$\lambda_1(u, a^I) \leqq m_1 \sum_{i \in I} a^i, \tag{40}$$

$$\lambda_0(u, a^I) \geqq m_0 \sum_{i \in I} a^i. \tag{41}$$

From (38), (40), and (41) we obtain

$$\frac{1}{m_1 \sum_{i\in I} a^i} \leqq \|M_I^{-1}(u, a^I)\| \leqq \frac{1}{m_0 \sum_{i\in I} a^i}. \tag{42}$$

Therefore it follows from (40) and (39) that

$$x^* G_{IJ}^*(u) M_I^{-1}(u, a^I) G_{IJ}(u) x \geqq \frac{|G_{IJ}(u)x|^2}{m_1 \sum_{i\in I} a^i}. \tag{43}$$

Since the rank of the matrix $G_{IJ}(u)$ is $|I|+|J|$, there exists an $m_2>0$ such that $|G_{IJ}(u)x| \geqq m_2|x|$ for each $u \in \bar{U}_{IJ}$. Also

$$|x|\,|G_{IJ}^*(u)M_I^{-1}(u, a^I)G_{IJ}(u)x| \geqq x^* G_{IJ}^*(u)M_I^{-1}(u, a^I)G_{IJ}(u)x \geqq \frac{m_2^2|x|^2}{m_1 \sum_{i\in I} a^i} \tag{44}$$

from which we obtain

$$|G_{IJ}^*(u)M_I^{-1}(u, a^I)G_{IJ}(u)x| \geqq \frac{m_2^2|x|}{m_1 \sum_{i\in I} a^i}. \tag{45}$$

We take $x=(G_{IJ}^* M_I^{-1} G_{IJ})^{-1} y$ and substitute into (45). This yields

$$|y| \geqq \frac{m_2^2 |(G_{IJ}^*(u)M_I^{-1}(u, a^I)G_{IJ}(u))^{-1} y|}{m_1 \sum_{i\in I} a^i}$$

for every y, i.e.,

$$\begin{aligned} &\sup_{y\neq 0} \frac{|(G_{IJ}^*(u)M_I^{-1}(u, a)G_{IJ}(u))^{-1}y|}{|y|} \\ &\quad = \|(G_{IJ}^*(u)M_I^{-1}(u, a^I)G_{IJ}(u))^{-1}\| \leqq \frac{m_1 \sum_{i\in I} a^i}{m_2^2} \end{aligned} \tag{46}$$

for every $u \in \bar{U}_{IJ}$.

Due to the equivalence of the Euclidean and operator norms, the assertion of the lemma follows directly from (42) and (46). □

LEMMA 17. *For every IJ admissible, $|I| \neq 0$, there exist positive constants k_3, k_4, k_5 such that for every $u \in \bar{U}_{IJ}$, $a^I \geqq 0$, $a^I \neq 0$*

$$|D^{-1}| \leqq k_1 \Big/ \sum_{i\in I} a^i + k_2 + k_3 \sum_{i\in I} a^I \tag{47}$$

where

$$D = \begin{pmatrix} S_J M_I & S_J(\partial g^I/\partial u)^* \\ G_{IJ}^* & 0 \end{pmatrix} \tag{48}$$

is an $(m+|I|)\times(m+|I|)$ matrix.

Proof. Let m_3 be a positive number such that

$$|G_{IJ}(u)| \geqq m_3 \tag{49}$$

for every $u \in \bar{U}_{IJ}$. We denote $h=(G_{IJ}M_I^{-1}G_{IJ})^{-1}$. Using the formula for the inverse of

a block matrix [5] and (49), (35), and (36) we obtain

$$\left|\begin{pmatrix} M_I & G_{IJ} \\ G_{IJ}^* & 0 \end{pmatrix}^{-1}\right| = \left|\begin{pmatrix} M_I^{-1}+M_I^{-1}G_{IJ}HG_{IJ}^*M_I^{-1} & -M_I^{-1}G_{IJ}H \\ HG_{IJ}^*M_I^{-1} & H \end{pmatrix}\right|$$

$$\leqq k_1 \Big/ \sum_{i\in I} a^i + k_1^2 m_3^2 k_2 \Big/ \sum_{i\in I} a^i + 2k_2 m_3 k_1 + k_2 \sum_{i\in I} a^i \tag{50}$$

$$= k_1' \Big/ \sum_{i\in I} a^i + k_2' + k_2 \sum_{i\in I} a^i.$$

We denote by N an $(m+|I|+|J|)\times(m+|I|+|J|)$ matrix that arises with the addition to D of some rows and columns by the formula

$$N = \begin{pmatrix} P_J M_I & 0 & P_J c^J \\ S_J M_I & S(\partial g^I/\partial u) & 0 \\ G_{IJ}^* & 0 & 0 \end{pmatrix}.$$

It is easy to see that

$$N = \begin{pmatrix} R_J & 0 \\ 0 & E \end{pmatrix}\begin{pmatrix} M_I & G_{IJ} \\ G_{IJ}^* & 0 \end{pmatrix} \tag{51}$$

where E is a $(|I|+|J|)\times(|I|+|J|)$ identity matrix. We denote

$$k_3' = \left|\begin{pmatrix} R_I & 0 \\ 0 & E \end{pmatrix}^{-1}\right| > 0. \tag{52}$$

Then, using (51) and (50), we obtain

$$|D^{-1}| \leqq |N^{-1}| \leqq k_3'\left(k_1' \Big/ \sum_{i\in I} a^i + k_2' + k_2 \sum_{i\in I} a^i\right).$$

Hence Lemma 17 is proved. □

LEMMA 18. *Let IJ be admissible such that $|I| \neq 0$. Let Z be a linear operator $Z: R^z \to R^n$. Then there are positive constants h_1, h_2, and h_3 such that*

$$\left|\frac{\partial w(Z\zeta)}{\partial \zeta}\right| \leqq \left(\frac{h_1}{|S_J B^* Z\zeta|} + h_2 + h_3|S_J B^* Z\zeta|\right)|S_J B^* Z| \tag{53}$$

for every $\zeta \in R^z$ such that $Z\zeta \in V_{IJ}$.

Proof. In Lemma 9 we have obtained $w(\psi)$ as a solution of the system (14)-(16). This system is equivalent to the system (29), (30), (15), (16) in which (29) is independent of the other equations. That is why by solving (30), (15), (16), we obtain the same solutions $w(\psi)$ as when we solve (14)-(16). Applying the formula for a derivative of a function given implicitly to $w(\psi)$ given by (30), (15), and (16), and using Lemmas 15 and 17, we obtain

$$\left|\frac{\partial w_{IJ}(Z\zeta)}{\partial \zeta}\right| \leqq |D^{-1}||S_J B^* Z| \leqq \left(k_1 \Big/ \sum_{i\in I} a^i + k_2 + k_3 \sum_{i\in I} a^i\right)|S_J B^* Z|$$

$$\leqq \left(\frac{k_1}{k_1'|S_J B^* Z\zeta|} + k_2 + k_3 k_1'|S_J B^* Z\zeta|\right)|S_J B^* Z|.$$

This proves the lemma. □

For every $i=1,\cdots,n-1$, J admissible, $|J|<m-1$, ($|J|=0$ as well) we define

$$V_{Ji}=\{\psi\in V_J / S_JB^*\psi=0,\cdots,S_JB^*A^{*i-1}\psi=0, S_JB^*A^{*i}\psi\neq 0\} \tag{54}$$

(for V_J see (6)).

LEMMA 19. *The family $\mathcal{P}$ of sets V_{Ji}, $i=1,\cdots,n-1$, $|J|<m-1$, J admissible, is a* CASA (*connected analytic submanifold which is a subanalytic set*) *stratification of W_Z, where by the* CASA *stratification we understand a stratification whose members are* CASA *sets.*

Proof. From the definition of V_{Ji} it is easy to see that every V_{Ji} is a CASA set. Now we prove that V_{Ji} form a partition of W_Z.

The sets V_J, $|J|<m-1$, form a partition of W_Z. Obviously $V_{Ji}\subseteq V_J$ and all the sets V_{Ji} are pairwise disjoint. It suffices to prove that V_{Ji}, $i=1,\cdots,n-1$ cover V_J.

Let there exist a $\psi\in V_J$ such that $S_JB^*A^{*i}\psi=0$ for every $i=1,\cdots,n-1$. Then from the normality condition it follows that there exists a $k>0$ such that $S_JB^*A^{*k}\psi\neq 0$ and $S_JB^*A^{*j}\psi=0$ for $j<k$. According to the Cayley-Hamilton theorem there exist constants $\alpha_1,\cdots,\alpha_n$ such that

$$\begin{aligned} S_JB^*A^{*k}\psi &= -S_JB^*(\alpha_1A^{*k-1}+\cdots+\alpha_nA^{*k-n})\psi \\ &= -\alpha_1S_JB^*A^{*k-1}\psi-\cdots-\alpha_nS_JB^*A^{*k-n}\psi=0, \end{aligned}$$

which contradicts our assumption. The sets V_{Ji} cover V_J. The stratification property follows directly from the definition of V_{Ji}. The lemma is proved. □

We denote by L_{Ji} the linear space given by the equations $S_JB^*A^{*j}\psi=0, j=1,\cdots,i-1$. Obviously V_{Ji} is an open subset of L_{Ji}. Let $\dim L_{Ji}$ be r_{Ji}, let Y_{Ji} be an $n\times r_{Ji}$ matrix of rank r_{Ji} such that the linear operator $Y_{Ji}: R^r\to L_{Ji}$ is a bijection.

LEMMA 20. *For every $\psi=Y_{Ji}\zeta\in V_{Ji}$ we have*

$$\lim_{t\to 0}\frac{|S_JB^*\mathcal{F}_tY_{Ji}|}{|S_JB^*\mathcal{F}_tY_{Ji}\zeta|}=\frac{|S_JB^*A^{*i}Y_{Ji}|}{|S_JB^*A^{*i}Y_{Ji}\zeta|} \tag{55}$$

where $\mathcal{F}_t=e^{-A^*t}$ is the flow of the adjoint equation.

Proof. It suffices to prove that

$$\lim_{t\to 0}\frac{|S_JB^*\mathcal{F}_tY_{Ji}|^2}{|S_JB^*\mathcal{F}_tY_{Ji}\zeta|^2}=\frac{|S_JB^*A^{*i}Y_{Ji}|^2}{|S_JB^*A^{*i}Y_{Ji}\zeta|^2}. \tag{56}$$

But (56) follows if we use L'Hôpital's rule $2i$ times and from the definition of V_{Ji}.

DEFINITION. Let K be a compact subset of an analytic submanifold M of R^n. Let $\mathcal{F}_t$ be a flow of an analytic vector field on R^n such that $\mathcal{F}_{[-T,0)}(K)\cap M=\emptyset$. Let g be a function defined on $\mathcal{F}_{[-T,0)}(K)\to R^m$ and let g be locally Lipschitz continuous on $\mathcal{F}_{[-T,0)}(K)$. We shall say that g can be extended as a Lipschitz continuous function on $\mathcal{F}_{[-T,0]}(K)$ along the flow $\mathcal{F}$ from the left, if for every $\tau\in(0,T)$ there exist an $L>0$ such that if $s\in[\tau,T]$ and $x_1,x_2\in K$; then

$$|g(\mathcal{F}_{t-s}(x_1))-g(\mathcal{F}_{t-s}(x_2))|\leqq L|\mathcal{F}_{-s}(x_1)-\mathcal{F}_{-s}(x_2)| \tag{57}$$

then every $t\in[0,s)$.

Remark 6. Analogously we can define the Lipschitz continuous extendability of a function g defined on $\mathcal{F}_{(0,T]}(K)$ to $\mathcal{F}_{[0,T]}(K)$ along the flow $\mathcal{F}$ from the right.

THEOREM 5. *Let IJ be admissible, $|I|\neq 0$, $i\in\{1,\cdots,n-1\}$. Let R be an analytic submanifold of V_{Ji} with the following property: for any compact subset K of R there exists a $T>0$ such that $\mathcal{F}_{[-T,0)}(K)\subset V_{IJ}$, $\mathcal{F}_t=e^{-A^*t}$. Then for every $\psi_0\in K$ there exists a compact neighbourhood O of ψ_0 in R such that the function $w_0(\psi)=w_{IJ}(\psi)/\mathcal{F}_{[-T,0)}(O)$ can be extended as a Lipschitz continuous function on $\mathcal{F}_{[-T,0]}(O)$ along the flow $\mathcal{F}_t$ from the left.*

Proof. Let $L = L_{Ji}$, $r = r_{Ji}$, $Y = Y_{Ji}$ be, respectively, the linear space, its dimension, and the matrix of the linear operator, corresponding to V_{Ji}. Let $\psi_0 \in R$ and let O_{ψ_0} be a compact neighbourhood of ψ_0 in R. For O_{ψ_0} there exists a compact neighbourhood $\tilde{O}_{\psi_0}$ of ψ_0 in L such that $\tilde{O}_{\psi_0} \cap R = O_{\psi_0}$. Because of the assumption of our theorem there is a $T > 0$ such that $\mathscr{F}_{[-T,0)}(O_{\psi_0}) \subset V_{IJ}$. The neighbourhoods O_{ψ_0} and $\tilde{O}_{\psi_0}$ could be chosen such that $\mathscr{F}_{[-T,0)}(\tilde{O}_{\psi_0}) \subset B_{IJ}$ where B_{IJ} is the neighbourhood of V_{IJ} in W_A (from Theorem 1) on which the analytic function $w_{IJ}(\psi)$ is defined.

Let $\zeta_0 \in R^r$ such that $\psi_0 = Y\zeta_0$. Let $O_{\psi_0} = YO_{\zeta_0}$ and $\tilde{O}_{\psi_0} = Y\tilde{O}_{\zeta_0}$. Let $\tau \in (0, T)$. We prove the existence of $L_1 > 0$ such that

$$\left|\frac{\partial w_{IJ}(\mathscr{F}_{t-s} Y\zeta)}{\partial \zeta}\right| \leqq L_1 \tag{58}$$

for all $t \in [0, s)$, $s \in [\tau, T)$, and $\zeta \in O_{\zeta_0}$.

According to Lemma 18 there exist positive constants h_1, h_2, h_3 such that

$$\left|\frac{\partial w_{IJ}(\mathscr{F}_{t-s} Y\zeta)}{\partial \zeta}\right| \leqq \left(\frac{h_1}{|S_J B^* \mathscr{F}_{t-s} Y\zeta|} + h_2 + h_3 |S_J B^* \mathscr{F}_{t-s} Y\zeta|\right) |S_J B^* \mathscr{F}_{t-s} Y|$$

for all $t \in [0, s)$, $s \in [\tau, T)$, and $\zeta \in O_{\psi_0}$.

From the definition of V_{Ji} it is easy to see that

$$\lim_{t \to s} |S_J B^* \mathscr{F}_{t-s} Y| = 0 \quad \text{and} \quad \lim_{t \to s} |S_J B^* \mathscr{F}_{t-s} Y\zeta| = 0.$$

By Lemma 20 we have

$$\lim_{t \to s} \frac{|S_J B^* \mathscr{F}_{t-s} Y|}{|S_J B^* \mathscr{F}_{t-s} Y\zeta|} = \frac{|S_J B^* A^{*i} Y|}{|S_J B^* A^{*i} Y\zeta|}.$$

Since $|S_J B^* A^{*i} Y\zeta|$ is bounded below (ζ lies in a compact set) the existence of an L_1 with property (58) is proved. Therefore the neighbourhoods O_{ψ_0} and O_{ζ_0} can be chosen such that

$$|w_{IJ}(\mathscr{F}_{t-s}(Y\zeta_1)) - w_{IJ}(\mathscr{F}_{t-s}(Y\zeta_2))| \leqq L_1 |\zeta_1 - \zeta_2| \tag{59}$$

for every $s \in [\tau, T)$, $t \in [0, s)$, and $\zeta_1, \zeta_2 \in O_{\zeta_0}$.

Further, the operator $\mathscr{F}_{-s} Y$ is linear and bounded below on the compact set O_{ζ_0}. Therefore there exists an $L_2 > 0$ such that for $s \in [\tau, T]$,

$$|\mathscr{F}_{-s}(Y\zeta_1) - \mathscr{F}_{-s}(Y\zeta_2)| \geqq L_2 |\zeta_1 - \zeta_2|. \tag{60}$$

Then from (59) and (60) we obtain

$$|w_{IJ}(\mathscr{F}_{t-s}(\psi_1) - w_{IJ}(\mathscr{F}_{t-s}(\psi_2))| \leqq L |\mathscr{F}_{-s}(\psi_1) - \mathscr{F}_{-s}(\psi_2)|,$$

where $L = L_1 / L_2$, for all $t \in [0, s)$, $s \in [\tau, T)$, and $\psi_1, \psi_2 \in O_{\psi_0}$. The theorem is proved. □

Remark 7. Analogously we can formulate and prove the modification of this theorem for the Lipschitz continuous extendability of the function $w_{IJ}(\psi)$ defined on $\mathscr{F}_{(0,T]}(O)$ to $\mathscr{F}_{[0,T]}(O)$ in the flow from the right.

8. The locally Lipschitz continuity of $F(\psi, T)$. Now, in addition to the analytic vector flow $\mathscr{F}_t = e^{-A^* t}$ of the adjoint equation on $R^n - \{0\}$, we shall consider the flow ϕ_t on $S^{n-1} = \{\psi \in R^n / |\psi| = 1\}$, which is a radial projection of $\mathscr{F}_t$. Let us denote by $\chi : R^n - \{0\} \to S^{n-1}$ the projection $\chi(\psi) = \psi |\psi|^{-1}$. Then $\phi_t(\chi(\psi)) = \chi(\mathscr{F}_t(\psi))$ for all t. For every $A \subseteq R^n - \{0\}$ such that $A \cup \{0\}$ is a cone, we shall denote by A^χ its radial

projection on S^{n-1}, that is, $A^X = \{\psi \in A/|\psi| = 1\}$. The set C defined by (25) can be written as

$$C = \{\psi \in S^{n-1}/\exists t \in R,\ \phi_t(\psi) \in W_Z^X\},$$

$$W_Z^X = \bigcup_{|J|<m-1} V_J^X.$$

By Lemma 2 every V_J^X is a subanalytic subset of S^{n-1}, and according to Corollary 3 we have dim $V_J^X < n-2$. Hence W_Z^X is a subanalytic set of dimension smaller than $n-2$. According to Lemma 19 the family $\mathcal{P}^X = \{V_{Ji}^X/J$ admissible, $i = 1, \cdots, n-1\}$ is a CASA stratification of W_Z^X with a finite number of members. Obviously the dimension of every member of $\mathcal{P}^X$ is smaller than $n-2$. For any $P \in \mathcal{P}^X$ and l natural, we define

$$C_{P,l} = \{\psi \in S^{n-1}/\exists t \in [-l, l] \text{ such that } \phi_t(\psi) \in P\}. \tag{61}$$

Evidently every $C_{P,l}$ is a subanalytic set the dimension of which is smaller than $n-1$ and

$$C = \bigcup_{P \in \mathcal{P}} \bigcup_{l=1}^{\infty} C_{P,l}.$$

Hence

$$H = F(C \times [0, \infty)) = \bigcup_{i=1}^{\infty} \bigcup_{P \in \mathcal{P}^X} \bigcup_{l=1}^{\infty} F(C_{P,l} \times [0, i]). \tag{62}$$

The dimension of the sets $C_{P,l} \times [0, i]$ is smaller than n. To prove $\mu(H) = 0$ it suffices to show (1) that for every $P \in \mathcal{P}^X$, l, i natural, there exists a locally finite partition of $C_{P,k} \times [0, k] \supset C_{P,l} \times [0, i]$, where $k = \max\{i, l\}$, whose components are analytic manifolds and (2) that the function $F(\psi, T)$ is locally Lipschitz continuous on every member of this partition.

THEOREM 6. *Let $P_0 \in \mathcal{P}^X$, l natural. Then there exists a* CASA *stratification $\mathcal{Y}$ of $C_{P_0,l}$ such that $F(\psi, T)$ is locally Lipschitz continuous on every $S \times [0, l]$, $S \in \varphi$.*

Proof. Let $\psi \in C_{P_0,l}$. Due to the normality condition the trajectory $\phi_t(\psi)$, $t \in [0, l]$ meets the set W_Z in at most a finite number of points. We define a function h from $C_{P_0,l}$ to the set of all natural numbers N that associates to each $\psi \in C_{P_0,l}$ the number of common points of the sets $\phi_{[0,l]}(\psi)$ and W_Z^X. From the definition of h it is easy to see that h is a subanalytic and locally bounded function (for the definitions see [10]). Then by [10] there exists a CASA stratification $\mathcal{Y}_1$ of $C_{P_0,l}$ compatible with $\mathcal{P}$ such that h is analytic and therefore constant on every member of $\mathcal{Y}_1$.

We choose an $S \in \mathcal{Y}_1$. Let $k \in N$ be the value of h that is constant on S. For given S, k consider finite sequences $i_1, \cdots, i_k$, where $i_j = (IJ_{i_j}^-, P_{i_j}, IJ_{i_j}^+)$, $j = 1, \cdots, k$, $IJ_{i_j}^+$, $IJ_{i_j}^-$ are admissible index sets such that if $IJ_{i_j}^- = J_{i_j}^-$, then $|J_{i_j}^-| = m$ and if $IJ_{i_j}^+ = J_{i_j}^+$, then $|J_{i_j}^+| = m$, $P_{i_j} \in \mathcal{P}$. We call such i_j admissible triples. For every sequence $i_1, \cdots, i_k$ of admissible triples we define the subset $S_{i_1 \cdots i_k}$ of S by

$$\psi \in S_{i_1 \cdots i_k} \Leftrightarrow (\psi \in S)$$

$$\wedge(\exists t_1)[(0 \leqq t_1 < l) \wedge (\phi_{t_1}(\psi) \in P_{i_1})$$

$$\wedge(\exists t_1^-)[(-\tfrac{1}{2} \leqq t_1^- < t_1) \wedge (\forall t)((t_1^- \leqq t < t_1) \Rightarrow (\phi_t(\psi) \in V_{IJ_{i_1}^-}))]$$

$$\wedge(\exists t_1^+)[(t_1 < t_1^+ \leqq l) \wedge (\forall t)((t_1 < t \leqq t_1^+) \Rightarrow (\phi_t(\psi) \in V_{IJ_{i_1}^+}))]]$$

$$\bigwedge(\exists t_k)[(t_{k-1}^+ < t_k \leqq l) \wedge (\phi_{t_k}(\psi) \in P_{i_k})$$
$$\bigwedge(\exists t_k^-)[(t_{k-1}^+ \leqq t_k^- < t_k) \wedge (\forall t)((t_k^- \leqq t < t_k) \Rightarrow (\phi_t(\psi) \in V_{IJ_i^-}))]$$
$$\bigwedge(\exists t_k^+)[(t_k \leqq t_k^+ \leqq l + \tfrac{1}{2}) \wedge (\forall t)((t_k < t \leqq t_k^+) \Rightarrow (\phi_t(\psi) \in V_{IJ_i^+}))]].$$

Roughly speaking, all the trajectories $\phi_t(\psi)$, $\psi \in S_{i_1 \cdots i_k}$, $t \in [0, l]$ meet the same subsets P_{i_j} of W_Z^X in the same order; they enter P_{i_j} from the same $V_{IJ_i^-}$ and go out of P_{i_j} to the same $V_{IJ_i^+}$.

The sets $S_{i_1 \cdots i_k}$ are subanalytic subsets of S. Further, for $(i_1, \cdots, i_k)$ running over all the ordered k-tuples of all admissible triples they constitute a finite partition of S. For every set $S_{i_1 \cdots i_k}$ we define a function $\sigma_{i_1 \cdots i_k} : S_{i_1 \cdots i_k} \to R^{3k}$ by the formula

$$\sigma_{i_1 \cdots i_k} = (t_1^-(\psi), t_1(\psi), t_1^+(\psi), \cdots, t_k^-(\psi), t_k(\psi), t_k^+(\psi))$$

where $t_j(\psi)$ are the points from the definition of $S_{i_1 \cdots i_k}$ such that $\phi_{t_j(\psi)}(\psi) \in P_{i_j}$ and

$$t_j^-(\psi) = \inf\{t / t_{j-1}^+(\psi) \leqq t \leqq t_j(\psi),\ \phi_t(\psi) \in V_{IJ_i^-}\},$$
$$t_j^+(\psi) = \sup\{t / t_j(\psi) \leqq t \leqq t_{j+1}^+(\psi),\ \phi_t(\psi) \in V_{IJ_i^+}\}$$

for all $j = 1, \cdots, k$, where $t_0^+(\psi) = -\frac{1}{2}$ and $t_{k+1}^-(\psi) = l + \frac{1}{2}$.

The function $\sigma_{i_1 \cdots i_k}$ is clearly subanalytic and bounded on $S_{i_1 \cdots i_k}$. Then, due to [10], there exists a CASA stratification of $S_{i_1 \cdots i_k}$ such that $\sigma_{i_1 \cdots i_k}$ is analytic on every member of this stratification. Since the function σ can be extended as a subanalytic and bounded function on the whole S^{n-1}, the stratifications of the sets $S_{i_1 \cdots i_k}$ are locally finite in S^{n-1} and therefore their union is locally finite in S^{n-1} as well. That is why there exists a CASA stratification $\mathscr{Y}$ of $C_{P_0, l}$ compatible with the partition $\mathscr{Y}_1$ and with $\mathscr{P}$ as well.

Let $S_0 \in \mathscr{Y}$. Then the following properties hold.

(a) There is a number $K(S_0) \in N$, a sequence of sets $P_1, \cdots, P_k \in \mathscr{P}$ and analytic functions $t_1(\psi), \cdots, t_k(\psi)$, $0 \leqq t_1(\psi) < \cdots < t_k(\psi) \leqq l$ such that $\mathscr{F}_{t_i(\psi)}(\psi) \in P_i$ and for any $t \neq t_i(\psi)$, $i = 1, \cdots, k$ we have $\mathscr{F}_t(\psi) \cap W_Z = \varnothing$.

(b) For any P_i, $i = 1, \cdots, k$, there are IJ_i^-, IJ_i^+ admissible such that if $IJ_i^+ = J_i^+$, then $|J_i^+| = m$ and there exist analytic functions $t_i^-(\psi)$, $t_i^+(\psi)$ such that $t_i^-(\psi) < t_i(\psi) < t_i^+(\psi)$ and $\mathscr{F}_t(\psi) \in V_{IJ^-}$ for every $t \in (t_i^-(\psi), t_i(\psi))$, and $\mathscr{F}_t(\psi) \in V_{IJ_i^+}$ for every $t \in (t_i(\psi), t_i^+(\psi))$ and $\psi \in S_0$, $i = 1, \cdots, k$.

From (a) and (b) it follows that

(c) The sets $R_i = \{\mathscr{F}_{t_i(\psi)}(\psi) / \psi \in S_0\}$, $i = 1, \cdots, k$ are analytic manifolds such that for any compact subset K of R_i there exist $T_i^- > 0$, $T_i^+ > 0$ such that

$$\mathscr{F}_{[-T_i^-, 0)}(K) \subset V_{IJ_i^-}, \qquad \mathscr{F}_{(0, T_i^+]}(K) \subset V_{IJ_i^+}.$$

We shall prove locally Lipschitz continuity of $F(\psi, T)$ on $S_0 \times [0, 1]$. Since the function $F(\psi, T)$ is continuous and Lipschitz continuous in T, it suffices to prove that for any $\psi_0 \in S_0$ there exists a neighbourhood O_{ψ_0} of ψ_0 such that $F(\psi, T)$ is Lipschitz continuous in $\psi \in O_{\psi_0}$ for any given T, $T \leqq l$.

Let $\psi_0 \in S_0$. Let O_{ψ_0} be a compact neighbourhood of ψ_0 in S_0. We denote $O_{\psi_0}^i = \mathscr{F}_{t_i(\psi)} O_{\psi_0}$, $i = 1, \cdots, k$. Obviously $O_{\psi_0}^i$ are compact neighbourhoods of R_i, $i = 1, \cdots, k$. Due to property (c) for any $O_{\psi_0}^i$ there exist $T_i^+ > 0$, $T_i^- > 0$ such that

$$\mathscr{F}_{[-T_i^-, 0)} O_{\psi_0}^i \subset V_{IJ_i^-} \quad \text{and} \quad \mathscr{F}_{(0, T_i^+]} O_{\psi_0}^i \subset V_{IJ_i^+}.$$

The neighbourhood O_{ψ_0} could be chosen suitably small such that there are τ_i^-, τ_i^+,

$$\tau_i^- < \min_{\psi \in O_{\psi_0}} t_i(\psi) \quad \text{and} \quad \mathscr{F}_{\tau_i^-}(O_{\psi_0}) \subset \mathscr{F}_{[-T_i^-, 0]}(O_{\psi_0}^i),$$
$$\max_{\psi \in O_{\psi_0}} t_i(\psi) < \tau_i^+ \quad \text{and} \quad \mathscr{F}_{\tau_i^-}(O_{\psi_0}) \subset \mathscr{F}_{(0, T_i^+]}(O_{\psi_0}^i)$$

for $i = 1, \cdots, k$.

Then by Theorem 5 there are $L_i^- > 0, L_i^+ > 0$ such that for any $s \in [\min_{\psi \in O_{\psi_0}} t_i(\psi) - \tau_i^-, T_i^-]$ or $s \in [\max_{\psi \in O_{\psi_0}} \tau_i^+ - t_i(\psi), T_i^+]$ and for each $\psi_1, \psi_2 \in O_{\psi_0}^i$ we have

$$|w(\mathscr{F}_{t-s}(\psi_1)) - w(\mathscr{F}_{t-s}(\psi_2))| \leqq L_i^- |\mathscr{F}_{-s}(\psi_1) - \mathscr{F}_{-s}(\psi_2)|, \tag{63}$$

or

$$|w(\mathscr{F}_{s-t}(\psi_1)) - w(\mathscr{F}_{s-t}(\psi_2))| \leqq L_i^+ |\mathscr{F}_s(\psi_1) - \mathscr{F}_s(\psi_2)| \tag{64}$$

for $t \in [0, s)$.

Let $\psi_1, \psi_2 \in O_{\psi_0}$ and let $T \leqq l$. We denote $t_0(\psi_1) = t_0(\psi_0) = 0$, $t_{k+1}(\psi_1) = t_{k+1}(\psi_1) = T$. Then

$$\begin{aligned} &|F(\psi_1, T) - F(\psi_2, T)| \\ &\leqq \sum_{j=1}^{k} \left| \int_{t_j(\psi_1)}^{t_{j+1}(\psi_1)} e^{-At} B v(\psi_1, t)\, dt - \int_{t_j(\psi_2)}^{t_{j+1}(\psi_2)} e^{-At} B v(\psi_2, t)\, dt \right|. \end{aligned} \tag{65}$$

The theorem will be proved if we find $L_j > 0$ for every $j = 0, \cdots, k$, such that

$$\left| \int_{t_j(\psi_1)}^{t_{j+1}(\psi_1)} e^{-At} B v(\psi_1, t)\, dt - \int_{t_j(\psi_2)}^{t_{j+1}(\psi_2)} e^{-At} B v(\psi_2, t)\, dt \right| \leqq L_j |\psi_1 - \psi_2|. \tag{66}$$

For simplicity, we prove the existence of the constant L_0 and the validity of (66) for $j = 0$. The relation (66) for other j's can be proved analogously.

Since the function $t_1(\psi)$ is analytic on O_{ψ_0}, there is a $k_1 > 0$ such that

$$|t_1(\psi_1) - t_1(\psi_2)| \leqq k_1 |\psi_1 - \psi_2|. \tag{67}$$

Obviously, there is a $k_2 > 0$ such that

$$|e^{-At} B| \leqq k_2 \tag{68}$$

for every $t \in [-l, l]$ and there is a k_3 such that

$$|v(\psi, t)| \leqq k_3 \tag{69}$$

for every ψ, t. Also, there is a k_4 such that

$$|\mathscr{F}_t| \leqq k_4 \tag{70}$$

for every $t \in [-l, l]$.

We assume that $0 < t_1(\psi_1) < t_1(\psi_2)$. Since $\mathscr{F}_{[0, \tau_1^-)}(O_{\psi_0}) \subset (R^n - \{0\}) - W_Z$, according to Lemma 14 there is a constant $k_5 > 0$ such that

$$|v(\psi_1, t) - v(\psi_2, t)| \leqq k_5 |\mathscr{F}_t(\psi_1) - \mathscr{F}_t(\psi_2)| \tag{71}$$

for each $t \in [0, \tau_i^-]$. Using (63), (67)-(71), and fundamental properties of integrals, we

can derive the following estimates:

$$\begin{aligned}
&\left|\int_0^{t_1(\psi_1)} e^{-At}Bv(\psi_1,t)\,dt-\int_0^{t_1(\psi_2)} e^{-At}Bv(\psi_2,t)\,dt\right| \\
&\quad\leqq\left|\int_0^{\tau_1^-}|e^{-At}B|\,|v(\psi_1,t)-v(\psi_2,t)|\,dt\right| \\
&\qquad+\left|\int_{\tau_1^-}^{t_1(\psi_1)} e^{-At}Bv(\psi_1,t)\,dt-\int_{\tau_1^-}^{t_1(\psi_2)} e^{-At}Bv(\psi_2,t)\,dt\right| \\
&\quad\leqq lk_2k_5k_4|\psi_1-\psi_2|+\left|\int_{\tau_1^-t_1(\psi_1)}^{0} e^{-A(q+t_1(\psi_1))}Bv(\psi_1,q+t_1(\psi_1))\,dq\right. \\
&\qquad\left.-\int_{\tau_1^-t_1(\psi_2)}^{0} e^{-A(q+t_1(\psi_2))}Bv(\psi_2,q+t_1(\psi_2))\,dq\right| \\
&\quad\leqq lk_2k_5k_4|\psi_1-\psi_2|+\left|\int_{\tau_1^-t_1(\psi_1)}^{0}|e^{-A(q+t_1(\psi_1))}B|\,|w(\mathscr{F}_{q+t_1(\psi_1)}\psi_1)-w(\mathscr{F}_{q+t_1(\psi_2)}\psi_2)|\,dq\right| \\
&\qquad+\left|\int_{\tau_1^-t_1(\psi_1)}^{0}|e^{-A(q+t_1(\psi_1))}-e^{-A(q+t_1(\psi_2))}|\,|v(\psi_2,q+t_1(\psi_2))|\,dq\right| \\
&\qquad+\left|\int_{\tau_1^-t_1(\psi_2)}^{\tau_1^-t_1(\psi_1)}|e^{-A(q+t_1(\psi_2))}B|\,|v(\psi_2,q+t_1(\psi_2))|\,dq\right| \\
&\quad\leqq lk_2k_4k_5|\psi_1-\psi_2|+lk_2L_1^-|\mathscr{F}_{-(t_1(\psi_1)-\tau_1^-)}\mathscr{F}_{t_1(\psi_1)}\psi_1-\mathscr{F}_{-(t_1(\psi_1)-\tau_1^-)}\mathscr{F}_{t_1(\psi_2)}\psi_2| \\
&\qquad+lk_3k_4|e^{-At_1(\psi_1)}-e^{-At_1(\psi_2)}|+k_1|\psi_1-\psi_2|k_2k_3.
\end{aligned}$$

Due to the analyticity of the e^{-At}, for $t\in[-l,l]$ there is a constant $k_6\geqq 0$ such that

$$|e^{-At_1(\psi_1)}-e^{-At_1(\psi_2)}|\leqq k_6|t_1(\psi_1)-t_1(\psi_2)|\leqq k_6k_1|\psi_1-\psi_2|.$$

Because of $\psi_1,\psi_2\in O_{\psi_0}$, where O_{ψ_0} is compact, there is a $k_7>0$ such that $|\psi|\leqq k_7$ for each $\psi\in O_{\psi_0}$. Then

$$\begin{aligned}
&|\mathscr{F}_{-(t_1(\psi_1)-\tau_1^-)}\mathscr{F}_{t_1(\psi_1)}\psi_1-\mathscr{F}_{-(t_1(\psi_1)-\tau_1^-)}\mathscr{F}_{t_1(\psi_2)}\psi_2| \\
&\quad\leqq k_4|\mathscr{F}_{t_1(\psi_1)}\psi_1-\mathscr{F}_{t_1(\psi_2)}\psi_2| \\
&\quad\leqq k_4(|\mathscr{F}_{t_1(\psi_1)}\psi_1-\mathscr{F}_{t_1(\psi_1)}\psi_2|+|\mathscr{F}_{t_1(\psi_1)}\psi_2-\mathscr{F}_{t_1(\psi_2)}\psi_2|) \\
&\quad\leqq k_4(k_4|\psi_1-\psi_2|+k_1k_6|\psi_1-\psi_2|).
\end{aligned}$$

This completes the proof.

COROLLARY 10. $\mu(H)=0$.

The proof follows directly from Theorem 6, (62), and the well-known fact that the image of the locally Lipschitz continuous function from an analytic submanifold of R^n of a dimension smaller than n to R^n is of the Lebesgue measure zero.

Acknowledgments. This paper, part of the author's doctoral thesis at Comenius University, was written under the guidance of P. Brunovský to whom the author expresses her thanks. The author also thanks the anonymous referee for suggestions and critical comments to the first version of this paper.

REFERENCES

[1] V. G. BOLTYANSKII, *Mathematical Methods of Optimal Control*, Nauka, Moscow, 1973.

[2] P. BRUNOVSKÝ, *Every normal linear system has a regular time optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81-100.

[3] ———, *Existence of regular synthesis for general control problems*, J. Differential Equations, 38 (1980), pp. 317-343.

[4] ———, *Regular synthesis for the linear quadratic optimal control problem with linear control constraints*, J. Differential Equations, 38 (1980), pp. 344-360.

[5] F. R. GANTMACHER, *Matrix Theory*, Nauka, Moscow, 1966.

[6] M. HALICKÁ, *Regular synthesis for a linear-convex optimal control problem with convex control constraint*, Math. Slovaca, 37 (1987), pp. 89-110.

[7] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[8] S. LOJASIEWICZ, JR. AND H. J. SUSSMANN, *Some examples of reachable sets and optimal cost functions that fail to be subanalytic*, SIAM J. Control Optim., 23 (1985), pp. 584-598.

[9] E. B. LEE AND L. MARCUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[10] H. J. SUSSMANN, *Analytic stratification and control theory*, in Proc. Internat. Congress of Mathematicians, Helsinki, 1978, pp. 865-871.

[11] H. J. SUSSMANN, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations, 43 (1982), pp. 399-418.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 330-373, March 1989

006

EXACT CONTROLLABILITY OF THE EULER-BERNOULLI EQUATION WITH CONTROLS IN THE DIRICHLET AND NEUMANN BOUNDARY CONDITIONS: A NONCONSERVATIVE CASE*[1]

I. LASIECKA† AND R. TRIGGIANI†

Abstract. This paper considers the Euler-Bernoulli problem (1.1a-d) with boundary controls g_1, g_2 in the Dirichlet and Neumann boundary conditions, respectively. Several exact controllability results are shown, including the following. Problem (1.1a-d) is exactly controllable in an arbitrarily short time $T>0$ in the space (of maximal regularity) $H^{-1}(\Omega)\times V'$, V as in (1.4), (i) with boundary controls $g_1\in L^2(\Sigma)$, $g_2\equiv 0$ under some geometrical conditions on Ω; (ii) with boundary controls $g_1\in L^2(\Sigma)$ and $g_2\in L^2(0, T; H^{-1}(\Gamma))$ without geometrical conditions on Ω. A direct approach is given, based on an operator model for problem (1.1a-d) and on multiplier techniques. An additional difficulty of the particular boundary conditions is due to the fact that, in the natural norms for the solution arising in the application of multiplier techniques, the corresponding homogeneous problem is *not* conservative. This difficulty is overcome by passing to an equivalent norm for the solution, with respect to which the homogeneous problem becomes conservative.

Key words. exact controllability, Euler-Bernoulli equation, boundary control

AMS(MOS) subject classifications. 35, 93

1. Introduction, statement of main results, and literature. Throughout this paper Ω is an open, bounded domain in R^n, $n\geqq 2$, with sufficiently smooth boundary $\partial\Omega=\Gamma$. In Ω, we consider the following nonhomogeneous problem for the Euler-Bernoulli equation in the solution $w(t, x)$:

$$w_{tt}+\Delta^2 w=0 \qquad \text{in } (0, T]\times\Omega\equiv Q, \tag{1.1a}$$

$$w(0,\circ)=w^0, \quad w_t(0,\circ)=w^1 \qquad \text{in } \Omega, \tag{1.1b}$$

$$w|_\Sigma\equiv g_1 \qquad \text{in } (0, T]\times\Gamma\equiv\Sigma, \tag{1.1c}$$

$$\left.\frac{\partial w}{\partial\nu}\right|_\Sigma\equiv g_2 \qquad \text{in } (0, T]\times\Gamma\equiv\Sigma, \tag{1.1d}$$

with control functions g_1, g_2 to be suitably selected below. Here, ν is the unit outward normal vector to Γ. In this paper we study the problem of *exact controllability* for the dynamics of (1.1a-d). As a matter of fact, the case $g_1\equiv 0$ and $g_2\in L^2(\Sigma)$ has already been studied by Lions in [L3, § 3], where he obtains exact controllability results in the space $L^2(\Omega)\times H^{-2}(\Omega)$ for T greater than some suitable $T_0>0$. Lions' results were then refined by Komornik [K1], who improved the estimate of T_0, and then by Zuazua [Z1], who showed that T can be taken arbitrarily small (as expected) by using an idea, first introduced in [BLR1], to prove a needed uniqueness result.

In the same Von Neumann Lecture [L3, Remark 3.5], Lions raises the question as to whether problem (1.1a-d) with boundary controls

$$g_1\in L^2(\Sigma), \qquad g_2\equiv 0 \tag{1.2}$$

(i.e., with purely Dirichlet control) is exactly controllable and, if so, in what space. In particular, Lions raises the question of characterizing his space F for problem (1.1a-d) subject to (1.2).

* Received by the editors October 28, 1987; accepted for publication (in revised form) May 9, 1988. This research was supported in part by the National Science Foundation grant NSF-DMS 8301668 and in part by the Air Force Office of Scientific Research grant AFOSR-84-0365.

† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22901.

[1] An announcement of the present paper with a brief sketch of the proofs is given in [LT6].

The main aim of the present paper is to provide affirmative answers to the above (and related) questions. To give a proper foundation for our subsequent analysis on exact controllability, we begin by stating regularity results for problem (1.1a-d) in the cases of interest. To this end, we introduce the following spaces:

$$X \equiv H^{-1}(\Omega) \times V', \tag{1.3}$$

$$V \equiv \left\{ f \in H^3(\Omega) : f|_\Gamma = \left. \frac{\partial f}{\partial \nu} \right|_\Gamma = 0 \right\}. \tag{1.4}$$

The space X in (1.3) can be likewise identified as

$$X \equiv [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]' \tag{1.5}$$

(with equivalent norms), where A is the positive self-adjoint operator defined by $Af = \Delta^2 f$, $\mathscr{D}(A) = H^4(\Omega) \cap H_0^2(\Omega)$ (see (2.2) below). In (1.5), the symbol $'$ denotes duality of $\mathscr{D}(A^{1/4})$ and $\mathscr{D}(A^{3/4})$, respectively, with respect to the $L_2(\Omega)$-topology. The norms are given by

$$\|x\|_{\mathscr{D}(A^\alpha)} \equiv \|A^\alpha x\|_{L^2(\Omega)}, \qquad \|x\|_{[\mathscr{D}(A^\beta)]'} = \|A^{-\beta} x\|_{L^2(\Omega)}, \tag{1.6}$$

where $\alpha, \beta \geqq 0$.

THEOREM 1.0 (Regularity). *Consider problem* (1.1a-d) *subject to*

$$\{w^0, w^1\} \in X, \quad g_1 \in L^2(\Sigma), \quad g_2 \in L^2(0, T; H^{-1}(\Gamma)).$$

Then the map

$$\{w^0, w^1, g_1, g_2\} \to \{w(t), w_t(t)\} \in C([0, T]; X)$$

is continuous for any $0 < T < \infty$, *where the space* X *is defined by* (1.3)-(1.5).

The proof of Theorem 1.0 follows by applying a transposition argument to recent results of Lions [L2] combined with cosine-sine operator theory on the initial data. Details are omitted. See also Remark 4.1 and (5.16) below.

Remark 1.1. For the purposes of the subsequent Theorem 1.2, we note that with, say, $w^0 = w^1 = g_2 \equiv 0$, the corresponding map

$$g_1 \to \{w(t), w_t(t)\}$$

is *not* continuous $H_0^1(0, T; L^2(\Gamma)) \to C([0, T]; Y)$, where (with equivalent norms):

$$Y \equiv H_0^1(\Omega) \times H^{-1}(\Omega) \equiv \mathscr{D}(A^{1/4}) \times [\mathscr{D}(A^{1/4})]'. \tag{1.7}$$

Instead, $g_1 \to \{w(T), w_t(T)\}$ is continuous into Y, by virtue of the smoothing due to the condition $g_1(T) = 0$. See more specifically Remark 3.1.

With the regularity theorem at hand, we can now state our main exact controllability results. They will be listed in the order in which they are proved, even though a given result may be extended by a subsequent one (e.g., Theorem 1.1 is improved in Theorem 1.4, etc.).

THEOREM 1.1. *Assume that there exists a point* $x_0 \in R^n$ *such that*

$$(x - x_0) \cdot \nu \geqq \text{const} = \gamma > 0 \quad \text{on } \Gamma. \tag{1.8}^2$$

---

[2] It is observed in [L7] that our proof here can be generalized to include the case $\gamma = 0$. This is done by using the estimate $\int_\Sigma |\nabla_\sigma(\Delta\phi)|^2 \, d\Sigma \leqq c_T \|\{\phi^0, \phi^1\}\|_Z^2$, $Z = \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})$ from [L2], in the analysis of the left-hand side of identity (2.24), carried out in step (vii) of the proof of Lemma 2.2, or in the proof of Theorem 4.5. The authors now have a different proof of the above estimate on $\nabla_\sigma(\Delta\phi)$.

Then there exists $T_0 > 0$ (*which can be explicitly estimated from the proof of Lemma* 2.2) *such that if* $T > T_0$, *then, given any pair of initial data* $\{w^0, w^1\} \in X$ (*see* (1.3)-(1.5)), *there exists a boundary control* $g_1 \in L^2(0, T; L^2(\Gamma))$ *such that the corresponding solution of problem* (1.1a-d), (1.2) *satisfies*

$$w(T, \cdot) = w_t(T, \cdot) = 0, \qquad \begin{vmatrix} w \\ w_t \end{vmatrix} \in C([0, T]; X).$$

Remark 1.2. Actually, the proof of Theorem 1.1 extends with no extra effort to domains Ω for which there exists a vector field $h(x) = [h_1(x), \cdots, h_n(x)]$ defined by

$$h_i(x) = \sum_{j=1}^{n} a_{ij}(x_j - x_{0,j}) \quad \text{for some } x_0 = [x_{0,1}, \cdots, x_{0,n}] \in R^n, \tag{1.9}$$

where the coefficients $\{a_{ij}\}$ satisfy (1.12) below, such that

$$h(x) \cdot \nu(x) \geqq \text{const} = \gamma > 0 \quad \text{on } \Gamma \tag{1.10}$$

as in (1.8).

Set

$$H(x) \equiv \begin{vmatrix} \partial h_1/\partial x_1 & \cdots & \partial h_1/\partial x_n \\ \vdots & & \vdots \\ \partial h_n/\partial x_1 & & \partial h_n/\partial x_n \end{vmatrix} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} \equiv H \tag{1.11}$$

and require that the symmetric matrix

$$H + H^* = \begin{vmatrix} 2a_{11} & \cdots & a_{1n} + a_{n1} \\ \vdots & & \vdots \\ a_{n1} + a_{1n} & \cdots & 2a_{nn} \end{vmatrix} \tag{1.12}$$

be strictly positive definite:

$$H + H^* \geqq \rho I \quad \text{for some } \rho > 0, \tag{1.13}$$

so that

$$\int_\Omega H(x)v(x) \cdot v(x)\, d\Omega \geqq \rho \int_\Omega |v(x)|^2_{R^n}\, d\Omega \quad \text{for all } v(x) \in [L^2(\Omega)]^n. \tag{1.14}$$

Moreover, the definition (1.9) of h implies (in the notation of § 6 below (see (6.4)))

$$4G_h \equiv \max_{\bar{\Omega}} |\nabla(\text{div } h)| = 0. \tag{1.15}$$

Subject to the standing assumption (1.10), the proof of Theorem 1.1 in § 2 applies equally well for linear vector fields $h(x)$, as in (1.9), which satisfy the positivity condition (1.13) (as well as (1.15)). See also the proof in § 6 for the more general situation of Theorem 1.5, or of Theorem 6.2; these results allow domains Ω for which there is a general vector field $h(x) \in C^2(\bar{\Omega})$ satisfying (1.10), (1.14) and another condition to take care of $G_h \neq 0$.

The next result considers a smoother control g_1 and consequently a smoother target space in part (i). In part (ii), it interpolates between Theorem 1.1 and Theorem 1.2(i).

THEOREM 1.2. (i) *Under condition* (1.8) *of Theorem* 1.1, *there exists* $T_0 > 0$ (*which can be explicitly estimated from the proof*) *such that if* $T > T_0$, *then given any pair of initial data* $\{w^0, w^1\} \in Y$ (*see* (1.7)), *there exists a boundary control* $g_1 \in H_0^1(0, T; L^2(\Gamma))$ *such that the corresponding solution to problem* (1.1a-d) *with such* g_1 *and* $g_2 \equiv 0$ *satisfies* $w(T, \cdot) = w_t(T, \cdot) = 0$.

(ii) *Moreover, for T sufficiently large, exact controllability of problem* (1.1a-d) *in the sense described by the preceding statement is equivalent to exact controllability of problem* (1.1a-d) *in the sense described by Theorem* 1.1, *and the simultaneous characterization of these two notions is given by* (2.11) *below. As a result, under assumption* (1.8), *an interpolation result between Theorem* 1.1 *and Theorem* 1.2 *is available and is described in detail in Corollary* 3.3(ii) *below.*

The next result manages to *dispense altogether with the geometrical condition* (1.8) imposed on the (smooth) domain Ω, at the price of inserting an additional control function g_2 in the Neumann boundary condition (1.1d).

Also, exact controllability is achieved in an arbitrarily short time.

THEOREM 1.3. *For any* $T > 0$, *given any pair of initial data* $\{w^0, w^1\} \in X$ (*see* (1.3)-(1.5)) *there exist boundary controls*

$$g_1 \in L^2(0, T; L^2(\Gamma)) \quad \text{and} \quad g_2 \in L^2(0, T; H^{-1}(\Gamma))$$

such that the corresponding solution of problem (1.1a-d) *satisfies*

$$w(T, \cdot) = w_t(T, \cdot) = 0, \qquad \begin{vmatrix} w \\ w_t \end{vmatrix} \in C([0, T]; X).$$

A direct extension of Theorem 1.3 to the case where the second control g_2 (with the same regularity) acts only on a suitable portion of the boundary is provided in Theorem 4.5, which likewise does not require geometrical conditions on Ω.

The proof of Theorem 1.3, suitably modified and complemented, will allow us to obtain the following improved version of Theorem 1.1.

THEOREM 1.4. *For the case* $g_1 \in L^2(\Sigma)$ *and* $g_2 \equiv 0$, *Theorem* 1.1 *admits a stronger conclusion in the sense that the time T of exact controllability stated there* (*universal, i.e., independent of the initial conditions*) *may be taken arbitrarily small.*

Remark 1.3. Consider the following homogeneous problem:

$$\phi_{tt} + \Delta^2 \phi \equiv 0 \qquad \text{in } Q, \tag{1.16a}$$

$$\phi|_{t=0} = \phi^0, \quad \phi_t|_{t=0} = \phi^1 \qquad \text{in } \Omega, \tag{1.16b}$$

$$\phi|_\Sigma \equiv 0 \qquad \text{in } \Sigma, \tag{1.16c}$$

$$\left.\frac{\partial \phi}{\partial \nu}\right|_\Sigma \equiv 0 \qquad \text{in } \Sigma. \tag{1.16d}$$

The proof of Theorem 1.1 shows that exact controllability of (1.1a-d)-(1.2) on the space X is *equivalent* to the following inequality: There exists C'_T such that

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial \nu}\right)^2 d\Sigma \geq C'_T \|\{\phi^0, \phi^1\}\|^2_{\mathcal{D}(A^{3/4}) \times \mathcal{D}(A^{1/4})}; \tag{1.17}$$

see the (backward in time) problem (2.9a-d) in Lemma 2.1. Lemma 2.2 shows then that this inequality holds true under (1.8) for $T > T_0$, where $C'_T = c(T - T_0)$ (or, more generally, for domains Ω that admit a linear field $h(x)$ as in (1.9) that satisfies (1.10) as well as (1.13). Moreover, Lemma 5.5 shows that T_0 can be taken arbitrarily small.

On the other hand, the opposite inequality:

$$\int_0^T \int_\Gamma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \leqq C_T \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{1.18}$$

is always true for all $0 < T < \infty$, as it follows by transposition from Lions' results [L2]. Hence for $T > T_0$, where T_0 can be taken arbitrarily small, and for Ω subject to condition (1.8) (or as in Remark 1.2), we can define a norm

$$\|\{\phi^0, \phi^1\}\|^2_F \equiv \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \tag{1.19}$$

on the space F which is, therefore, $F \equiv \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4}) \equiv V \times H_0^1(\Omega)$. Such a norm (1.12) is equivalent to the norm

$$\|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \equiv \|A^{3/4}\phi^0\|^2_{L_2(\Omega)} + \|A^{1/4}\phi^1\|^2_{L_2(\Omega)}. \tag{1.20}$$

This answers a question raised by Lions [L3, Remark 3.5].

Remark 1.4. Inequality (1.17) (to be proved in Lemma 2.2 for T_0 finite and in Lemma 5.5 for T_0 arbitrarily small for domains Ω for which there exists a vector field $h(x)$, as in (1.9), that satisfies (1.13), in particular a radial vector field) establishes a fortiori an apparently new uniqueness theorem under condition (1.10) (respectively, (1.8)): If ϕ solves (1.16a) and moreover the three boundary conditions

$$\phi|_\Sigma \equiv 0, \quad \left.\frac{\partial\phi}{\partial\nu}\right|_\Sigma \equiv 0, \quad \left.\frac{\partial(\Delta\phi)}{\partial\nu}\right|_\Sigma \equiv 0, \quad T > 0 \text{ arbitrary},$$

then $\phi \equiv 0$ in Q.

By contrast, a standard uniqueness theorem (Holmgren-John) (see [H2, Thm. 5.33, p. 129]), to be crucially invoked in the proof of Theorem 1.3; see (4.39) in the proof of Lemma 4.4, is instead: If ϕ solves (1.16a) and moreover the four boundary conditions

$$\phi|_\Sigma \equiv 0, \quad \left.\frac{\partial\phi}{\partial\nu}\right|_\Sigma \equiv 0, \quad \Delta\phi|_\Sigma \equiv 0, \quad \left.\frac{\partial(\Delta\phi)}{\partial\nu}\right|_\Sigma \equiv 0$$

for $T > 0$ arbitrary, then $\phi \equiv 0$ in Q.

The result in Theorem 1.1 admits the following generalization, whose proof will be given in § 6. It requires, in addition to the reasoning in the proof of Theorem 1.1, a nontrivial extra argument, which is presented first (in a slightly different context) in the proof of § 4 to establish that with two control functions g_1 and g_2, the time T for exact controllability can be taken arbitrarily small.

THEOREM 1.5. *The conclusion of Theorem* 1.1 *on exact controllability on* $X = [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$ *(see* (1.5)*) over any* $[0, T]$, $T > T_u$ *defined by* (1.24) *below holds true more generally for (smooth) domains* Ω *possessing the following geometrical properties: there exists a (general) vector field* $h(x) \in C^2(\bar{\Omega})$ *such that*

(1.21) (i_1) $h(x) \cdot \nu(x) \geqq \text{const} = \gamma > 0$ *on* Γ;
(i_2) *If we set*

$$H(x) \equiv \begin{vmatrix} \partial h_1/\partial x_1 & \cdots & \partial h_1/\partial x_n \\ \vdots & & \vdots \\ \partial h_n/\partial x_1 & \cdots & \partial h_n/\partial x_n \end{vmatrix}, \tag{1.22}$$

then

(1.23)

$$\int_\Omega H(x)v(x)\cdot v(x)\,d\Omega \geqq \rho \int_\Omega |v(x)|^2_{R^n}\,d\Omega \quad \textit{for some } \rho>0 \textit{ and all } v(x)\in[L^2(\Omega)]^n.$$

(*A sufficient checkable condition for* (1.23) *to hold true is that the symmetric matrix* $H(x)+H^*(x)$ *be uniformly strictly positive definite on* $\bar{\Omega}$.)

(ii) *Moreover,* Ω *satisfies the following uniqueness property*:

(1.24)

$$\phi_{tt}+\Delta^2\phi\equiv 0 \quad \textit{in } Q, \quad \phi|_\Sigma\equiv 0, \quad \left.\frac{\partial\phi}{\partial\nu}\right|_\Sigma\equiv 0, \quad \left.\frac{\partial(\Delta\phi)}{\partial\nu}\right|_\Sigma\equiv 0 \quad \textit{for } 0<t\leqq T_u$$

$$\textit{implies } \phi\equiv 0 \quad \textit{in } Q.$$

(*As shown in Remark* 1.4, *such a uniqueness property is satisfied for any* $T_u>0$ *by domains* Ω *for which there is a radial vector field* $x-x_0$ *satisfying* (1.8), *or more generally, for which there is a linear vector field as in* (1.9) *satisfying* (1.13) (*i.e.*, (1.23)) *as well as* (1.10) = (1.21).)

Remark 1.5 (on the smoothness of Γ). The proofs given below require the existence of a dense set of initial data for which the solutions of the corresponding homogeneous problem (1.16a-d) possess the regularity required to carry out the actual computations in the multiplier methods of Lemma 2.2, Proposition 4.2, etc.

The proofs of the preceding results will be given in the next sections according to the following strategy. In §§ 2 and 3 we provide the proofs of Theorems 1.1 and 1.2 in the case where $g_2=0$ and for T sufficiently large, respectively, in the special but important case in which the assumed vector field $h(x)$ is radial, or as in Remark 1.2. These sections form the core of proofs in the general case. The additional arguments required to prove, say, Theorem 1.1 for a general vector field as in Theorem 1.5 are presented in § 6. When we pass from a radial to a general vector field, the main nontrivial extra difficulty consists in the need to "absorb" lower-order interior terms on Q by appropriate boundary terms on Σ. This is accomplished in § 6 by reasoning in the same conceptual way (but with different details) as in the preceding §§ 4 and 5 in proving Theorems 1.3 and 1.4 for T arbitrarily small. (Reasoning of the same conceptual type, again with different technical details depending on the circumstances, is also used for wave equations [L3], [LT3], and [T2], and crucially exploited in [L8] and [L9].)

The exact controllability theorems (Theorems 1.1, 1.3, and 1.4) in the space of regularity X (see (1.3)-(1.5)) have an important implication in the quadratic cost problem over an infinite interval (regulator problem) corresponding to the dynamics (1.1a-d): Minimize

(1.25)

$$J(g_1,g_2,w) \equiv \int_0^\infty \left\{ \|w(t)\|^2_{[\mathscr{D}(A^{1/4})]'} + \|w_t(t)\|^2_{[\mathscr{D}(A^{3/4})]'} + \|g_1(t)\|^2_{L^2(\Gamma)} + \|g_2(t)\|^2_{H^{-1}(\Gamma)} \right\} dt$$

over all $\{g_1,g_2\}\in L^2(0,\infty;L^2(\Gamma))\times L^2(0,\infty;H^{-1}(\Gamma))\equiv U$. Indeed, Theorems 1.1 and 1.3 guarantee a fortiori that the corresponding finite-cost condition be fulfilled: For each $\{w^0,w^1\}\in X$ there is some $\{\bar{g}_1,\bar{g}_2\}\in U$ such that for the corresponding solution $\{\bar{w}(t),\bar{w}_t(t)\}$ we have $J(\bar{g}_1,\bar{g}_2,\bar{w})<\infty$, and indeed g_2 may be taken to be zero.

On the other hand, problem (1.1a-d) fits the abstract model considered in [FLT1], which offers a rather comprehensive study of the regulator problem and corresponding

algebraic Riccati equation (see Appendix 2 of [FLT1]). Thus we have the following theorem.

THEOREM 1.6. *The regulator theory of* [FLT1] *with cost function* (1.25) *applies to problem* (1.1a-d), (1.2) *under assumption* (1.8) *of Theorem* 1.1. *It provides, among other things, a boundary feedback* $g_1^0(t) = -B^*P[w^0(t), w_t^0(t)]$ *(in the notation of* [FLT1]*, we have* $u^0(t) = \{g_1^0(t), 0\}$*, where* $[g_1^0(t), \{w^0(t), w_t^0(t)\}]$ *is the unique optimal pair in problem* (1.25) *and P is a Riccati operator). Such feedback inserted in* (1.1c) *produces exponential decay in the uniform norm X of the corresponding feedback system. More specifically we have from* [FLT1] *(see* (2.1a-c) *and* (2.4) *below*):

$$\begin{aligned} g_1^0(t) &= -G_1^* A^{-1/2}[P\{w(t), w_t(t)\}]_2 \\ &= -G_1^* A A^{-3/2}[P\{w(t), w_t(t)\}]_2 \\ &= \frac{\partial \Delta}{\partial \nu} A^{-3/2}[P\{w(t), w_t(t)\}]_2, \end{aligned}$$

where $[y]_2$ *means the second component* y_2 *of the vector* $y = [y_1, y_2]$*. Alternatively, the regulator theory of* [FLT1] *applies with both* g_1 *and* g_2*, but with no geometrical conditions (Theorem* 1.3).

Notation. Unless otherwise specified, the norms $\| \ \|_\Omega$, $\| \ \|_\Gamma, \cdots$ and the inner products $(,)_\Omega$, $(,)_\Gamma, \cdots$ are all L^2 over the specified domain Ω, Γ, Q, Σ, etc.

Orientation. Our strategy in this paper consists, in short, of two main points (as in [T2], [LT3], and [LT8]):

(i) An operator approach to identifying an *equivalent* condition for exact controllability, in terms of the corresponding homogeneous partial differential problem with solution ϕ (such a condition is a bound from below of suitable traces of ϕ on Σ in terms of the interior norm of the initial data on Ω).

(ii) A multiplier technique for proving the condition in (i). Here we use the multipliers $h \cdot \nabla(\Delta\phi)$ and $\Delta\phi$ div h, where $h(x)$ is a smooth vector field on $\bar{\Omega}$. Instead, in [L3, § 3] for $g_1 = 0$ and $g_2 \in L^2(\Sigma)$, the multipliers $h \cdot \nabla\phi$ and ϕ, h a radial field, are employed (these are the same multipliers that were successful in the treatment of regularity and exact controllability questions for wave equations [L1]-[L6], [LLT1], [H1], [LT3], and [T2]). An additional difficulty of the particular boundary conditions in this paper is due to the fact that the natural norms arising in the application of the multiplier technique are

$$\left\{\int_\Omega |\nabla(\Delta\phi)|^2 \, d\Omega\right\}^{1/2} \quad \text{and} \quad \left\{\int_\Omega |\nabla\phi_t|^2 \, d\Omega\right\}^{1/2}$$

and in these norms the homogeneous problem in ϕ is *nonconservative.* (This is a novel feature over past literature on these problems.) This difficulty is overcome by realizing that the above norms are *equivalent* to the norms

$$\|A^{3/4}\phi\|_{L^2(\Omega)} \quad \text{and} \quad \|A^{1/4}\phi_t\|_{L^2(\Omega)}$$

with respect to which the homogeneous problem in ϕ becomes conservative. Once exact controllability is established, the simple argument of Appendix D provides the minimal norm steering control. The corresponding uniform stabilization problem is studied in [BT1].

2. Proof of Theorem 1.1. The radial vector field case.

Step 0. In line with the authors' approach to time-invariant problems with second-order differential operators in the space variables [LT1]-[LT5] and [T1]-[T3] we introduce an explicit input $g_1 \to$ solution $[w, w_t]$ map. To this end, we first define an

operator G_1 (Green map) by

$$G_1 g = y \Leftrightarrow \begin{cases} \Delta^2 y = 0 & \text{in } \Omega, \quad (2.1a) \\ y|_\Gamma = g, & (2.1b) \\ \partial y/\partial \nu|_\Gamma = 0, & (2.1c) \end{cases}$$

which is continuous $L^2(\Gamma) \to L^2(\Omega)$ (in fact, $L^2(\Gamma) \to H^{1/2}(\Omega)$ [LM1, Vol. I, pp. 188-189]). Next, we define the operator A: $L^2(\Omega) \supset \mathscr{D}(A) \to L^2(\Omega)$ by ([F1, p. 101])

$$\begin{aligned} &Af = \Delta^2 f, \\ &\mathscr{D}(A) = \left\{ f \in L^2(\Omega) : \Delta^2 f \in L^2(\Omega), f|_\Gamma = \left.\frac{\partial f}{\partial \nu}\right|_\Gamma = 0 \right\} = H^4(\Omega) \cap H_0^2(\Omega). \end{aligned} \quad (2.2)$$

The operator A is positive self-adjoint: $(Af, f)_\Omega = \|\Delta f\|_\Omega^2$ for $f \in \mathscr{D}(A)$, by Green's second theorem. Thus, $-A$ generates a strongly continuous (s.c.) cosine operator $C(t)$ on $L^2(\Omega)$, $t \in R$, with $S(t)z = \int_0^t C(\tau)z\, d\tau$, $z \in L^2(\Omega)$. Then, as in the authors' references above, the solution to problem (1.1a-d) with $w^0 = w^1 = 0$ at time T can be written in operator form as

$$\begin{vmatrix} w(T, t=0; w^0=0, w^1=0) \\ w_t(T, t=0; w^0=0, w^1=0) \end{vmatrix} = \mathscr{L}_{1T} g_1 = \begin{vmatrix} A \int_0^T S(T-t) G_1 g_1(t)\, dt \\ A \int_0^T C(T-t) G_1 g_1(t)\, dt \end{vmatrix}. \quad (2.3)$$

The following lemma in the style of [T1]-[T3] and [LT2] will be needed.

LEMMA 2.0. *Let G_1^* denote the continuous operator $L^2(\Omega) \to L^2(\Gamma)$, which is the adjoint of G_1: $(G_1 g, v)_\Omega = (g, G_1^* v)_\Gamma$, $g \in L^2(\Gamma)$, $v \in L^2(\Omega)$. Then*

$$G_1^* Af = \left.\frac{\partial(\Delta f)}{\partial \nu}\right|_\Gamma, \qquad f \in \mathscr{D}(A). \quad (2.4)$$

Proof. For $f \in \mathscr{D}(A)$ and $g \in L^2(\Omega)$ we compute by Green's second theorem applied twice:

$$\begin{aligned} (G_1^* Af, g)_\Gamma &= (Af, G_1 g)_\Omega = (\Delta(\Delta f), G_1 g)_\Omega \\ &= (\Delta f, \Delta(G_1 g))_\Omega + \left(\frac{\partial(\Delta f)}{\partial \nu}, G_1 g\right)_\Gamma - \left(\Delta f, \frac{\partial(G_1 g)}{\partial \nu}\right)_\Gamma \quad \text{(by (2.1c))} \\ &= (f, \Delta^2(G_1 g))_\Omega + \left(\frac{\partial f}{\partial \nu}, \Delta(G_1 g)\right)_\Gamma - \left(f, \frac{\partial(\Delta G_1 g)}{\partial \nu}\right)_\Gamma \\ &\qquad + \left(\frac{\partial(\Delta f)}{\partial \nu}, g\right)_\Gamma \quad \text{(by 2.1a) and (2.2))} \\ &= \left(\frac{\partial(\Delta f)}{\partial \nu}, g\right)_\Gamma. \end{aligned}$$

□

Step 1. The (regularity) Theorem 1.0 gives us that $\mathscr{L}_{1T}$ is continuous $L^2(\Sigma) \to X \equiv [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$. By time reversibility of problem (1.1a-d), exact controllability of problem (1.1a)-(1.1d) on the space X over $[0, T]$ is equivalent to

$$\mathscr{L}_{1T} : L^2(\Sigma) \xrightarrow{\text{ONTO}} X \equiv [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]' \quad (2.5)$$

which in turn is equivalent [TL1, p. 235] to the condition that the Hilbert space adjoint $\mathscr{L}_{1T}^*$ has continuous inverse: $X \to L^2(\Sigma)$ ("continuous observability" in the terminology of [DR1]); i.e., there exists $C_T > 0$ such that

$$\left\| \mathscr{L}_{1T}^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right\|_{L^2(\Sigma)} \geqq C_T \|\{z_1, z_2\}\|_{[\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'} \tag{2.6}$$

for all $z = \{z_1, z_2\} \in X$, where for $g \in L^2(\Sigma)$

$$(\mathscr{L}_{1T} g, z)_X = (g, \mathscr{L}_{1T}^* z)_{L^2(\Sigma)}. \tag{2.7}$$

Step 2. An equivalent partial differential equation characterization of inequality (2.6) is given by the following lemma.

LEMMA 2.1. (i) *For* $z = \{z_1, z_2\} \in X \equiv [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$, *we have*

$$\left(\mathscr{L}_{1T}^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right)(t) = \frac{\partial}{\partial \nu} (\Delta \phi(t)) \quad \text{on } \Sigma, \tag{2.8}$$

where $\phi(t) \equiv \phi(t; \phi^0, \phi^1)$ *is the solution of the following backward problem*:

$$\phi_{tt} + \Delta^2 \phi \equiv 0 \qquad \text{in } Q, \tag{2.9a}$$

$$\phi|_{t=T} = \phi^0, \quad \phi_t|_{t=T} = \phi^1 \quad \text{in } \Omega, \tag{2.9b}$$

$$\phi|_\Sigma \equiv 0 \qquad \text{in } \Sigma, \tag{2.9c}$$

$$\left. \frac{\partial \phi}{\partial \nu} \right|_\Sigma \equiv 0 \qquad \text{in } \Sigma, \tag{2.9d}$$

with

$$\phi^0 = A^{-3/2} z_2, \quad \phi^1 = -A^{-1/2} z_1 \quad \text{in } \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4}). \tag{2.10}$$

(ii) *For any* $0 < T < \infty$, (2.5) *is equivalent to the following*:
There exists $C_T' > 0$ *such that*

$$\int_\Sigma \left(\frac{\partial(\Delta \phi)}{\partial \nu} \right)^2 d\Sigma \geqq C_T' \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})} \tag{2.11}$$

for all $\{\phi^0, \phi^1\} \in \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})$.

Proof. (i) As in [LT2]-[LT3] and [T1]-[T3] we compute from (2.3), (2.5), and (1.6)

$$\begin{aligned}
\left(\mathscr{L}_{1T} g_1, \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right)_X &= \left(\begin{vmatrix} A \int_0^T S(T-t) G_1 g_1(t)\, dt \\ A \int_0^T C(T-t) G_1 g_1(t)\, dt \end{vmatrix}, \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right)_{[\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'} \\
&= \left(A \int_0^T S(T-t) G_1 g_1(t)\, dt, A^{-1/2} z_1 \right)_\Omega \\
&\qquad + \left(\int_0^T C(T-t) G_1 g_1(t)\, dt, A^{-1/2} z_2 \right)_\Omega \\
&= (g_1, G_1^* A[S(T-t) A^{-1/2} z_1 + C(T-t) A^{-3/2} z_2])_\Sigma \\
&= (g_1, G_1^* A[C(t-T) A^{-3/2} z_2 + S(t-T)(-A^{-1/2} z_1)])_\Sigma,
\end{aligned} \tag{2.12}$$

where in the last step we have used $C(\cdot)$ even while $S(\cdot)$ is odd. Thus from (2.7) and (2.12) we deduce

$$\left(\mathscr{L}_{1T}^{*}\begin{vmatrix} z_1 \\ z_2 \end{vmatrix}\right)(t) = G_1^{*}A[C(t-T)A^{-3/2}z_2 + S(t-T)(-A^{-1/2}z_1)] \quad \text{on } \Sigma. \tag{2.13}$$

But the solution of problem (2.9a-d), (2.10) is precisely

$$\phi(t) = C(t-T)\phi^0 + S(t-T)\phi^1 \tag{2.14}$$

and by (2.4)

$$\frac{\partial(\Delta\phi(t))}{\partial\nu} = G_1^{*}A[C(t-T)\phi^0 + S(t-T)\phi^1] \quad \text{on } \Sigma. \tag{2.15}$$

Comparing (2.13) with (2.15) (with initial data as in (2.10)) yields the desired conclusion (2.8). Then part (i) immediately implies part (ii) via (2.8) used in (2.6), with initial data

$$\begin{aligned} z_1 = -A^{1/2}\phi^1 \in [\mathscr{D}(A^{1/4})]' \quad &\text{so that } A^{-1/4}z_1 = -A^{1/4}\phi^1 \in L^2(\Omega) \\ &\text{and} \quad \phi^1 \in \mathscr{D}(A^{1/4}), \end{aligned} \tag{2.16}$$

$$\begin{aligned} z_2 = A^{3/2}\phi^0 \in [\mathscr{D}(A^{3/4})]' \quad &\text{so that } A^{-3/4}z_2 = A^{3/4}\phi^0 \in L^2(\Omega) \\ &\text{and} \quad \phi^0 \in \mathscr{D}(A^{3/4}). \end{aligned} \tag{2.17}$$

Hence,

$$\begin{aligned} &\|\{z_1 = -A^{1/2}\phi^1,\ z_2 = A^{3/2}\phi^0\}\|_{[\mathscr{D}(A^{1/4})]'\times[\mathscr{D}(A^{3/4})]'} \\ &\quad \text{is equivalent to } \|\{\phi^0, \phi^1\}\|_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}. \end{aligned} \tag{2.18}$$

The proof of Lemma 2.1 is complete. □

Step 3. It remains to show if or when (2.11) holds true. The following lemma is the key technical issue of the exact controllability problem for (1.1a-d).

LEMMA 2.2. *Under the same assumptions as in Theorem* 1.1, *there is* $T_0 > 0$ *as in Theorem* 1.1 *such that if* $T > T_0$, *then* (2.11) *holds true with* $C'_T = c(T - T_0)$:

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \geqq c(T-T_0)\|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{2.19}$$

where, by time reversal in (2.9a-d), *we may take* ϕ *to be the solution of the homogeneous problem*

$$\phi_{tt} + \Delta^2\phi \equiv 0 \qquad \text{in } Q, \tag{2.20a}$$

$$\phi(0,\cdot) = \phi^0 \in \mathscr{D}(A^{3/4}), \qquad \phi_t(0,\cdot) = \phi^1 \in \mathscr{D}(A^{1/4}) \quad \text{in } \Omega, \tag{2.20b}$$

$$\phi|_\Sigma \equiv 0 \qquad \text{in } \Sigma, \tag{2.20c}$$

$$\left.\frac{\partial\phi}{\partial\nu}\right|_\Sigma \equiv 0 \qquad \text{in } \Sigma. \tag{2.20d}$$

Remark 2.1. As pointed out in Remark 1.3, (1.18), the inequality opposite to (2.19), i.e.,

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \leqq CT\|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{2.21}$$

always holds true (with Γ sufficiently smooth as in Remark 1.4) for all $0 < T < \infty$, as

a consequence of Lions' results [L2] followed by a transposition argument.

Proof of Lemma 2.2.

Step (i). Let $h(x)=x-x_0$, the radial vector field assumed in the statement of Theorem 1.1. With reference to Remark 1.4 we multiply (2.20a) by the multiplier $h\cdot\nabla(\Delta\phi)$ and integrate over $\int_0^T\int_\Omega dQ$. We obtain (see Appendix A)

$$\begin{aligned}&\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\,h\cdot\nabla(\Delta\phi)\,d\Sigma-\frac{1}{2}\int_\Sigma|\nabla(\Delta\phi)|^2h\cdot\nu\,d\Sigma\\&=\int_Q|\nabla\phi_t|^2+|\nabla(\Delta\phi)|^2\,dQ+\frac{n}{2}\int_Q\{|\nabla\phi_t|^2-|\nabla(\Delta\phi)|^2\}\,dQ\\&\quad-[(\phi_t,h\cdot\nabla(\Delta\phi))_\Omega]_0^T\end{aligned}\tag{2.22}$$

after using the boundary conditions (2.20c-d).

Step (ii). We estimate the second integral on the right of (2.22). We multiply (2.20a) by $\Delta\phi$ and obtain (Appendix B)

$$\int_Q\{|\nabla\phi_t|^2-|\nabla(\Delta\phi)|^2\}\,dQ=\left[\int_\Omega\nabla\phi\cdot\nabla\phi_t\,d\Omega\right]_0^T-\int_\Sigma\frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi\,d\Sigma\tag{2.23}$$

after using the boundary conditions (2.20c-d).

Step (iii). Thus, inserting (2.23) into (2.22) results in

$$\begin{aligned}&\int_\Sigma\frac{\partial(\Delta\phi)}{\partial\nu}\,h\cdot\nabla(\Delta\phi)\,d\Sigma+\frac{n}{2}\int_\Sigma\frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi\,d\Sigma-\frac{1}{2}\int_\Sigma|\nabla(\Delta\phi)|^2h\cdot\nu\,d\Sigma\\&=\int_0^T\left\{\int_\Omega|\nabla\phi_t|^2+|\nabla(\Delta\phi)|^2\,d\Omega\right\}dt+\beta_{0,T},\end{aligned}\tag{2.24}$$

where $\beta_{0,T}$ (boundary terms at $t=T$ and $t=0$) is

$$\beta_{0,T}=\frac{n}{2}\left[\int_\Omega\nabla\phi\cdot\nabla\phi_t\,d\Omega\right]_0^T-[(\phi_t,h\cdot\nabla(\Delta\phi))_\Omega]_0^T.\tag{2.25}$$

Step (iv). We now use the standard fact that with A the positive self-adjoint operator defined by (2.2), the operator

$$\mathscr{A}=\begin{vmatrix}0 & I\\ -A & 0\end{vmatrix},\qquad \mathscr{D}(\mathscr{A})=\mathscr{D}(A)\times\mathscr{D}(A^{1/2}),\tag{2.26}$$

which describes the dynamics of (2.20),

$$\frac{d}{dt}\begin{vmatrix}\phi\\ \phi_t\end{vmatrix}=\mathscr{A}\begin{vmatrix}\phi\\ \phi_t\end{vmatrix},\tag{2.27}$$

is the generator of an s.c. *unitary* group on the space $\mathscr{D}(A^{1/2})\times L^2(\Omega)$. Accordingly, A with domain $\mathscr{D}(A^{5/4})\times\mathscr{D}(A^{3/4})$ generates likewise an s.c. *unitary* group on the space

$$Z=\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})\tag{2.28}$$

so that from (2.27),

$$\begin{aligned}\left\|\begin{vmatrix}\phi(t)\\ \phi_t(t)\end{vmatrix}\right\|_Z^2&\equiv\|\phi(t)\|^2_{\mathscr{D}(A^{3/4})}+\|\phi_t(t)\|^2_{\mathscr{D}(A^{1/4})}\\&\equiv\|A^{3/4}\phi(t)\|_\Omega^2+\|A^{1/4}\phi_t(t)\|_\Omega^2\quad\text{(from (1.6))}\\&=\left\|e^{\mathscr{A}t}\begin{vmatrix}\phi^0\\ \phi^1\end{vmatrix}\right\|_Z^2\equiv\left\|\begin{vmatrix}\phi^0\\ \phi^1\end{vmatrix}\right\|_Z^2\\&=\|A^{3/4}\phi^0\|_\Omega^2+\|A^{1/4}\phi^1\|_\Omega^2,\qquad t\in R,\end{aligned}\tag{2.29}$$

the norm-preserving identity to be crucially exploited below. Identity (2.29) is also obtained by multiplying problem (2.20a-d), rewritten abstractly as $\phi_{tt}+A\phi=0$, by $A^{1/2}\phi_t$ and integrating over $L^2(Q)$.

Step (v). Crucial to the utilization of (2.29) in analyzing (2.24) is the following lemma.

LEMMA 2.3. (i) *With reference to the positive self-adjoint operator A in* (2.2), *we have*

$$\mathscr{D}(A^{3/4})=\left\{f\in H^3(\Omega): f|_\Gamma=\left.\frac{\partial f}{\partial \nu}\right|_\Gamma=0\right\}, \tag{2.30}$$

$$\mathscr{D}(A^{1/4})=\{f\in H^1(\Omega): f|_\Gamma=0\}=H_0^1(\Omega) \tag{2.31}$$

the identifications being set theoretically and topologically, with equivalent norms. In particular, we have the following:

(ii) *For* $f\in\mathscr{D}(A^{3/4})$, *the norms*

$$\|f\|_{\mathscr{D}(A^{3/4})}=\|A^{3/4}f\|_\Omega \quad \textit{and} \quad \left\{\int_\Omega |\nabla(\Delta f)|^2\,d\Omega\right\}^{1/2} \tag{2.32}$$

are equivalent.

(iii) *Similarly, for* $f\in\mathscr{D}(A^{1/4})$, *the norms*

$$\|f\|_{\mathscr{D}(A^{1/4})}=\|A^{1/4}f\|_\Omega \quad \textit{and} \quad \left\{\int_\Omega |\nabla f|^2\,d\Omega\right\}^{1/2} \tag{2.33}$$

are equivalent.

(iv) *Parts* (ii) *and* (iii) *apply in particular to the solution* ϕ *of problem* (2.20a-d) *or* (2.27), *for which we have that the norm*

$$\begin{aligned}\left\|\begin{matrix}\phi(t)\\ \phi_t(t)\end{matrix}\right\|_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} &\equiv\{\|A^{3/4}\phi(t)\|_\Omega^2+\|A^{1/4}\phi_t(t)\|_\Omega^2\}^{1/2}\\ &\equiv\{\|A^{3/4}\phi^0\|_\Omega^2+\|A^{1/4}\phi^1\|_\Omega^2\}^{1/2},\end{aligned} \tag{2.34}$$

which is time-invariant by (2.29), *and the norm*

$$\left\{\int_\Omega |\nabla(\Delta\phi)|^2\,d\Omega+\int_\Omega |\nabla\phi_t|^2\,d\Omega\right\}^{1/2}=\{\|\nabla(\Delta\phi)\|_\Omega^2+\|\nabla\phi_t\|_\Omega^2\}^{1/2} \tag{2.35}$$

are equivalent.

Proof of Lemma 2.3. See Appendix C for the proof. □

Remark 2.2. Multiplying problem (2.20a-d) by $\Delta\phi_t$ and applying Green's theorems yields, as usual,

$$\frac{1}{2}\frac{\partial}{\partial t}\left\{\int_\Omega |\nabla\phi_t|^2+|\nabla(\Delta\phi)|^2\,d\Omega\right\}=\int_\Gamma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi_t\,d\Gamma. \tag{2.36}$$

Thus, we cannot claim by (2.36) that the norm occurring naturally in the multiplier method leading to identity (2.24) is time-invariant, as the boundary term at the right of (2.36) not only cannot be claimed to vanish, but moreover presents a delicate issue as to its well-posedness. (In fact, from the initial data in (2.20b) we deduce only that

$$\phi(t)=C(t)\phi^0+S(t)\phi^1\in C([0,T];\mathscr{D}(A^{3/4})),$$

$$\phi_t(t)=C(t)\phi^1-AS(t)\phi^0\in C([0,T];\mathscr{D}(A^{1/4})),$$

so that the expression in the brackets { } in (2.36) is continuous in time.) On the other hand, the equivalent norm (2.34) is time-invariant. Thus, in analyzing the terms in identity (2.24) we shall always refer to the time-invariant norm (2.34).

Step (vi). We now analyze the term $\beta_{0,T}$ in (2.25) by referring to the norm (2.28) of Z for $\{\phi, \phi_t\}$, as mentioned in Remark 2.2. For the first term in (2.25) we have by the Schwarz inequality

$$\left|\left[\int_\Omega \nabla\phi\cdot\nabla\phi_t\,d\Omega\right]_0^T\right| \leqq |||\nabla\phi(T)|||_\Omega |||\nabla\phi_t(T)|||_\Omega + |||\nabla\phi^0|||_\Omega |||\nabla\phi^1|||_\Omega$$

$$\leqq \tfrac{1}{2}\{|||\nabla\phi(T)|||_\Omega^2 + |||\nabla\phi_t(T)|||_\Omega^2 + |||\nabla\phi^0|||_\Omega^2 + |||\nabla\phi^1|||_\Omega^2\}$$

(using the norm-equivalence (2.33))

$$\leqq \frac{C}{2}\{\|A^{1/4}\phi(T)\|_\Omega^2 + \|A^{1/4}\phi_t(T)\|_\Omega^2 + \|A^{1/4}\phi^0\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2\}$$

$$\leqq \frac{C}{2}\{\|A^{3/4}\phi(T)\|_\Omega^2 + \|A^{1/4}\phi_t(T)\|_\Omega^2 + \|A^{3/4}\phi^0\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2\}.$$

Thus

$$\left|\left[\int_\Omega \nabla\phi\cdot\nabla\phi_t\,d\Omega\right]_0^T\right| \leqq C\{\|A^{3/4}\phi^0\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2\} \quad \forall T\in R, \tag{2.37}$$

where in the last step we have used the time invariance (2.29). Similarly for the second term in (2.25) we obtain by using the Poincaré inequality:

$$\int_\Omega \psi^2\,d\Omega \leqq C_p^2\int_\Omega |\nabla\psi|^2\,d\Omega, \quad \psi\in H_0^1(\Omega), \quad C_p = \text{Poincaré constant}$$

on $\phi^1\in H_0^1(\Omega)$ and on ϕ_t (legitimate by $\phi_{t|\Sigma}=0$, in view of (2.20c)) and $2M_h = \max_{\bar\Omega}|h|$:

$$|[\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T| \leqq 2M_h\{\|\phi_t(T)\|_\Omega |||\nabla(\Delta\phi(T))|||_\Omega + \|\phi^1\|_\Omega |||\nabla(\Delta\phi^0)|||_\Omega\}$$

$$\leqq 2M_hC_p\{|||\nabla\phi_t(T)|||_\Omega |||\nabla(\Delta\phi(T))|||_\Omega + |||\nabla\phi^1|||\nabla(\Delta\phi^0)|||_\Omega\}$$

$$\leqq M_hC_p\{|||\nabla\phi_t(T)|||_\Omega^2 + |||\nabla(\Delta\phi(T))|||_\Omega^2 + |||\nabla\phi^1|||_\Omega^2 + |||\nabla(\Delta\phi^0)|||_\Omega^2\}$$

(using the norm-equivalence (2.32) and (2.33))

$$\leqq M_hCC_p\{\|A^{1/4}\phi_t(T)\|_\Omega^2 + \|A^{3/4}\phi(T)\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2 + \|A^{3/4}\phi^0\|_\Omega^2\}.$$

Thus by the time invariance (2.29) we obtain

$$|[(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T| \leqq 2M_hCC_p\{\|A^{1/4}\phi^1\|_\Omega^2 + \|A^{3/4}\phi^0\|_\Omega^2\} \quad \forall T\in R, \tag{2.38}$$

where C_p is the Poincaré constant. Thus from (2.25), (2.37), and (2.38), we conclude that for all $0<T<\infty$,

$$|\beta_{0,T}| \leqq \text{const}_{h,n}\{\|A^{3/4}\phi^0\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2\} = \text{const}_{h,n}\|\{\phi^0,\phi^1\}\|_Z^2, \tag{2.39}$$

with Z as in (2.28).

Remark 2.3. The present method of estimating $\beta_{0,T}$ leads to (2.39) with a *finite* constant in front of $\|\{\phi^0,\phi^1\}\|_Z^2$. This fact is responsible for obtaining, in Theorem 1.1, exact controllability only for T greater than a finite time $T_0>0$ (see (2.50)–(2.52) below).

In Lemma 4.3 of § 4, we shall instead estimate $\beta_{0,T}$ in a different way and obtain in (4.21) an *arbitrarily small constant* in front of $\|\{\phi^0, \phi^1\}\|_Z^2$, at the price, however, of introducing two new terms in the interior Q. Via an argument that ultimately rests on a uniqueness theorem, these two additional terms over Q will then be "absorbed" for an arbitrary $T>0$ into the desired boundary terms over Σ that occur when a second control g_2 is also active in (1.1d). This way, exact controllability on an arbitrarily short time is achieved in Theorem 1.3 and in Proposition 5.1, both cases having g_2 active in (1.1d). However, this latter result can be improved by using it, in a kind of bootstrap argument, in combination with Lemma 2.2 for T large, and with a regularity result. This will be done in § 5.2, which is based on a recent idea of [BLR1]. Here we finally obtain T arbitrarily small, also with $g_1 \in L^2(\Sigma)$ and $g_2 \equiv 0$, as in Theorem 1.4.

Step (vii). Returning to the left-hand side (LHS) of identity (2.24), we compute with $2C_h = \max_\Gamma |h|$ and for any $\varepsilon > 0$:

First term.

$$\left| \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h \cdot \nabla(\Delta\phi)\, d\Sigma \right| \leqq C_h \int_\Sigma \left\{ \frac{1}{\varepsilon} \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 + \varepsilon |\nabla(\Delta\phi)|^2 \right\} d\Sigma; \tag{2.40}$$

Second term.

$$\begin{aligned} \left| \frac{n}{2} \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} \Delta\phi\, d\Sigma \right| &\leqq \frac{n}{4} \int_\Sigma \left\{ \frac{1}{\varepsilon} \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 + \varepsilon |\Delta\phi|^2 \right\} d\Sigma \\ &= \frac{n}{4\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 d\Sigma + \frac{n\varepsilon}{4} \|\Delta\phi\|^2_{L^2(0,T;L^2(\Gamma))} \\ &\leqq \frac{n}{4\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 d\Sigma + n\varepsilon C \|\phi\|^2_{L^2(0,T;\mathscr{D}(A^{3/4}))}, \end{aligned} \tag{2.41}$$

where in the last step we have used trace theory followed by Lemma 2.3, (2.30) for ϕ (which satisfies (2.20c–d)).

$$\|\Delta\phi\|_\Gamma \leqq c \|\Delta\phi\|_{H^1(\Omega)} \leqq C \|\phi\|_{H^3(\Omega)} \leqq C \|\phi\|_{\mathscr{D}(A^{3/4})}; \tag{2.42}$$

Third term. Using assumption (1.8) on the radial vector field $h(x)$ we have

$$-\frac{1}{2} \int_\Sigma |\nabla(\Delta\phi)|^2 h \cdot \nu\, d\Sigma \leqq -\frac{\gamma}{2} \int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma. \tag{2.43}$$

Summing up (2.40), (2.41), and (2.43) we obtain

$$\begin{aligned} &\left(\frac{C_h}{\varepsilon} + \frac{n}{4\varepsilon} \right) \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 d\Sigma + \left(C_h\varepsilon - \frac{\gamma}{2} \right) \int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma \\ &\quad + n\varepsilon c \|\phi\|^2_{L^2(0,T;\mathscr{D}(A^{3/4}))} \geqq \text{LHS of (2.24)}. \end{aligned} \tag{2.44}$$

Now, choosing $\varepsilon > 0$ sufficiently small to make $(C_h\varepsilon - \gamma/2) < 0$, we drop the second integral in (2.44) and finally obtain

$$C_{1,nh\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu} \right)^2 d\Sigma + n\varepsilon c \int_0^T \|A^{3/4}\phi\|^2_\Omega\, dt \geqq \text{LHS of (2.24)}. \tag{2.45}$$

Step (viii). We now work on the right-hand side (RHS) of identity (2.24), where we use the equivalence of Lemma 2.3(iii) (and time-invariance (2.29)), along with the bound (2.39). We obtain

$$\begin{aligned} \text{RHS of (2.24)} &\geqq C_2 \int_0^T \|A^{3/4}\phi\|^2_\Omega + \|A^{1/4}\phi_t\|^2\, dt - C_{h,n} \|\{\phi^0, \phi^1\}\|^2_Z \\ &= C_2 T \|\{\phi^0, \phi^1\}\|^2_Z - C_{h,n} \|\{\phi^0, \phi^1\}\|^2_Z. \end{aligned} \tag{2.46}$$

Step (ix). Combining (2.45) and (2.46), we obtain, using again the time-invariance (2.29) and $Z = \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})$ as in (2.28),

$$C_{1,nh\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + n\varepsilon c \int_0^T \|A^{3/4}\phi\|_\Omega^2 + \|A^{1/4}\phi_t\|_\Omega^2 \, dt$$

$$\equiv C_{1,nh\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + Tn\varepsilon c \|\{\phi^0, \phi^1\}\|_Z^2 \geqq \text{LHS of (2.24)} \tag{2.47}$$

$$= \text{RHS of (2.24)} \geqq C_2 T \|\{\phi^0, \phi^1\}\|_Z^2 - C_{hn} \|\{\phi^0, \phi^1\}\|_Z^2.$$

From here we obtain

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \geqq T\frac{(C_2 - n\varepsilon c)}{C_{1,nh\varepsilon}} \|\{\phi^0, \phi^1\}\|_Z^2 - \frac{C_{hn}}{C_{1nh\varepsilon}} \|\{\phi^0, \phi^1\}\|_Z^2 \tag{2.48}$$

and selecting $\varepsilon > 0$ suitably small so that

$$C'_{\varepsilon nh} = \frac{C_2 - n\varepsilon c}{C_{1,h\varepsilon n}} > 0, \tag{2.49}$$

we finally arrive at

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \geqq C'_{\varepsilon,nh}(T - T_0) \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})} \tag{2.50}$$

with

$$T_0 = \frac{C_{hn}}{C_{1nh\varepsilon}} \frac{1}{C'_{\varepsilon nh}} = \frac{C_{hn}}{C_2 + n\varepsilon c} > 0, \tag{2.51}$$

which is precisely (2.19). Lemma 2.2 is proved. □

Lemmas 2.1 and 2.2 along with (2.5) and (2.6) prove Theorem 1.1. □

Remark 2.4. For future use (in the proof of Theorem 1.4 in § 5), we note that the argument above yields the following inequality, which is more precise than (2.48):

$$C_{1,nh\varepsilon} \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \geqq (C_2 - n\varepsilon c) T \|\{\phi^0, \phi^1\}\|_Z^2 - |\beta_{0,T}|,$$
$$Z = \mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4}). \tag{2.52}$$

3. Proof of Theorem 1.2. We parallel and complement the proof of Theorem 1.1.

Step 1. As mentioned in Remark 1.1, the operator $\mathscr{L}_{1T}$ in (2.3) is continuous, $H_0^1(0, T; L^2(\Gamma)) \to H_0^1(\Omega) \times H^{-1}(\Omega)$ (see Remark 3.1), and the exact controllability requirement is

$$\mathscr{L}_{1T} : H_0^1(0, T; L^2(\Gamma)) \xrightarrow{\text{ONTO}} H_0^1(\Omega) \times H^{-1}(\Omega) \equiv \mathscr{D}(A^{1/4}) \times [\mathscr{D}(A^{1/4})]'. \tag{3.1}$$

But the $H_0^1(0, T)$-norm is equivalent to the gradient-norm. Then the condition—equivalent to (3.1)—that the corresponding Hilbert space adjoint $\mathscr{L}_{1T}^*$ of $\mathscr{L}_{1T}$ have continuous inverse can now be expressed: There exists $C_T > 0$ such that

$$\left\| \frac{d}{dt} \left(\mathscr{L}_{1T}^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right) \right\|_{L^2(\Sigma)} \geqq C_T \|\{z_1, z_2\}\|_{\mathscr{D}(A^{1/4}) \times [\mathscr{D}(A^{1/4})]'}, \tag{3.2}$$

where now for $g_1 \in H_0^1(0, T; L^2(\Gamma))$ and $z = [z_1, z_2] \in \mathcal{D}(A^{1/4}) \times [\mathcal{D}(A^{1/4})]'$ we have

$$(\mathcal{L}_{1T} g_1, z)_{\mathcal{D}(A^{1/4}) \times [\mathcal{D}(A^{1/4})]'} = (g_1, \mathcal{L}_{1T}^* z)_{H_0^1(0,T;L^2(\Gamma))} = \left(\frac{d}{dt} g_1, \quad \frac{d}{dt} \mathcal{L}_{1T}^* z \right)_{L^2(\Sigma)}$$

$$= \left(g_1, -\frac{d^2}{dt^2} \mathcal{L}_{1T}^* z \right)_{L^2(\Sigma)} \tag{3.3}$$

since g_1 vanishes at $t = 0$ and $t = T$.

Step 2. An equivalent partial differential equation characterization of (3.2) is given by the following lemma.

LEMMA 3.1. *With reference to* $\mathcal{L}_{1T}^*$ *in* (3.3) *we have for* $z = [z_1, z_2] \in \mathcal{D}(A^{1/4}) \times [\mathcal{D}(A^{1/4})]'$:

(i)

$$(\mathcal{L}_{1T}^* z)(t) = G_1^*[C(t-T)A^{-1/2} z_2 + S(t-T)(-A^{1/2} z_1)] + K_1 t + K_2 \in H_0^1(0, T; L^2(\Gamma)), \tag{3.4}$$

$$K_1 = K_{1T} = \frac{G_1^*}{T} \{[C(T) - I]A^{-1/2} z_2 + S(T)A^{1/2} z_1\}, \tag{3.5a}$$

$$K_2 = K_{2T} = -G_1^*[C(T)A^{-1/2} z_2 + S(T)A^{1/2} z_1], \tag{3.5b}$$

$$\frac{d}{dt}(\mathcal{L}_{1T}^* z)(t) = -\frac{\partial(\Delta\phi(t))}{\partial\nu} + K_{1T}, \tag{3.6}$$

where $\phi(t) = \phi(t, \phi^0, \phi^1)$ *is the solution of the following homogeneous problem, backward in time*

$$\phi_{tt} + \Delta^2 \phi = 0, \tag{3.7a}$$

$$\phi|_{t=T} = \phi^0, \quad \phi_t|_{t=T} = \phi^1, \tag{3.7b}$$

$$\phi|_\Sigma \equiv 0, \tag{3.7c}$$

$$\left.\frac{\partial\phi}{\partial\nu}\right|_\Sigma \equiv 0, \tag{3.7d}$$

with

$$\phi^0 = A^{-1/2} z_1, \qquad \phi^1 = A^{-1/2} z_2 \tag{3.8}$$

explicitly given by

$$\phi(t) = C(t-T)\phi^0 + S(t-T)\phi^1. \tag{3.9}$$

(ii) *For any* $T > 0$, (3.2) (*which characterizes exact controllability of problem* (1.1a–d) *with* $g_1 \in H_0^1(0, T; L^2(\Gamma))$ *and* $g_2 \equiv 0$ *on the space* $\mathcal{D}(A^{1/4}) \times [\mathcal{D}(A^{1/4})]'$ *over the interval* $[0, T]$) *is equivalent to saying: There exists* $C_T' > 0$ *such that*

$$\int_\Sigma \left[-\frac{\partial(\Delta\phi)}{\partial\nu} + K_{1T} \right]^2 d\Sigma \geqq C_T' \|\{\phi^0, \phi^1\}\|^2_{\mathcal{D}(A^{3/4}) \times \mathcal{D}(A^{1/4})}, \tag{3.10}$$

$$K_{1T} = \frac{G_1^*}{T} \{[C(T) - I]\phi^1 + AS(T)\phi^0\} = \frac{1}{T} \frac{\partial}{\partial\nu} \Delta\{[C(T) - I]A^{-1}\phi^1 + S(T)\phi^0\}. \tag{3.11}$$

Proof of Lemma 3.1. (i) With $g \in H_0^1(0, T; L^2(\Gamma))$ we compute from (2.3) and (1.6) by proceeding as in the proof of Lemma 2.1:

$$(\mathscr{L}_{1T} g_1, z)_{\mathscr{D}(A^{1/4})\times[\mathscr{D}(A^{1/4})]'} = \left(\int_0^T S(T-t)G_1 g_1(t)\, dt, A^{3/2} z_1\right)_{L^2(\Omega)}$$

$$+\left(\int_0^T C(T-t)G_1 g_1(t)\, dt, A^{1/2} z_2\right)_{L^2(\Omega)} \tag{3.12}$$

$$= \int_0^T (g_1(t), G_1^*[S(T-t)A^{3/2} z_1 + C(T-t)A^{1/2} z_2])_{L^2(\Gamma)}\, dt.$$

By comparing (3.3) with (3.12), we conclude that

$$\frac{-d^2}{dt^2}(\mathscr{L}_{1T}^* z)(t) = G_1^*[C(t-T)A^{1/2} z_2 + S(t-T)(-A^{3/2} z_1)], \tag{3.13}$$

since $C(\cdot)$ is even and $S(\cdot)$ is odd. Integrating in t, we find

$$\frac{d}{dt}(\mathscr{L}_{1T}^* z)(t) = -G_1^*[S(t-T)A^{1/2} z_2 - A^{-1}C(t-T)(-A^{3/2} z_1)] + K_1, \tag{3.14}$$

$$(\mathscr{L}_{1T}^* z)(t) = G_1^*[A^{-1}S(t-T)(-A^{3/2} z_1) + A^{-1}C(t-T)A^{1/2} z_2] + K_1 t + K_2. \tag{3.15}$$

By imposing that $(\mathscr{L}_{1T}^* z)(t)$ vanishes at $t=0$ and $t=T$, so that $\mathscr{L}_{1T}^* z \in H_0^1(0, T; L^2(\Gamma))$, we readily identify the operators K_1 and K_2 as in (3.5a-b) and then (3.15) becomes (3.4). For the purposes of (3.2), we now rewrite (3.14) as

$$\frac{d}{dt}(\mathscr{L}_{1T}^* z)(t) = -G_1^* A[C(t-T)A^{-1/2} z_1 + S(t-T)A^{-1/2} z_2] + K_{1T}$$

$$= -\frac{\partial(\Delta\phi(t))}{\partial\nu} + K_{1T}, \tag{3.16}$$

where in the last step we have used (2.4) and (3.9), (3.8), and part (i) is proved.

(ii) We first note that by (3.8)

$$\|\{z_1 = A^{1/2}\phi^0, z_2 = A^{1/2}\phi^1\}\|^2_{\mathscr{D}(A^{1/4})\times[\mathscr{D}(A^{1/4})]'}$$

$$= \|\{A^{3/4}\phi^0, A^{1/4}\phi^1\}\|^2_{L^2(\Omega)\times L^2(\Omega)} \tag{3.17}$$

$$= \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}.$$

Moreover, by virtue of (3.8) and (2.4), then (3.5a) becomes (3.11). Then (3.2) becomes (3.10) as desired, by use of (3.6), (3.11), and (3.17). The proof of Lemma 3.1 is complete. □

Step 3. The next lemma provides a sufficient condition for the exact controllability of problem (1.1a-d) considered in the present section for T arbitrary.

LEMMA 3.2. *A sufficient condition for* (3.10) *to hold is that there exists a constant* $C_T' > 0$ *such that*

$$\int_\Sigma \left(\frac{\partial(\Delta\phi(t))}{\partial\nu}\right)^2 d\Sigma \geqq C_T' \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} + 2T\|K_{1T}\|^2_{L^2(\Gamma)}, \tag{3.18}$$

where the term K_{1T} *is given by* (3.11) *and satisfies*

$$2\int_\Sigma |K_{1T}|^2\, d\Sigma = 2T\|K_{1T}\|^2_{L^2(\Gamma)} \leqq f(T)\|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}, \tag{3.19}$$

$$f(T)=4\frac{\|G_1^*\|^2}{T}\max\{\|[C(T)-I]A^{-1/4}\|^2,\|A^{1/4}S(T)\|^2\} \tag{3.20a}$$

$$\leqq\frac{C_f}{T},\qquad T>0. \tag{3.20b}$$

Proof. From $(a+b)^2=a^2+b^2+2ab\geqq a^2+b^2-\varepsilon a^2-1/\varepsilon b^2$ we obtain $(a+b)^2\geqq\frac{1}{2}a^2-b^2$ by selecting $\varepsilon=\frac{1}{2}$. Thus,

$$\begin{aligned}\int_\Sigma\left[-\frac{\partial(\Delta\phi)}{\partial\nu}+K_{1T}\right]^2 d\Sigma&\geqq\frac{1}{2}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma-\int_\Sigma|K_{1T}|^2\,d\Sigma\\&=\frac{1}{2}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma-T\|K_{1T}\|^2_{L^2(\Gamma)}.\end{aligned} \tag{3.21}$$

Thus, recalling (3.17), we see from (3.21) that (3.18) implies (3.10) as desired. To show (3.19) we rewrite K_{1T} in (3.11) accordingly as

$$K_{1T}=\frac{G_1^*}{T}[(C(T)-I)A^{-1/4}A^{1/4}\phi^1+A^{1/4}S(T)A^{3/4}\phi^0]. \tag{3.22}$$

□

Step 4. We now show that, under assumption (1.8), the analysis of § 2 guarantees that the sufficient condition (3.18) is fulfilled for T sufficiently large. Indeed, Lemma 2.2 yields (2.19), which we rewrite here for convenience

$$\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma\geqq c(T-T_0)\|\{\phi^0,\phi^1\}\|_Z^2,\qquad Z\equiv\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4}), \tag{3.23}$$

with c, T_0 identified in (2.50), (2.51). Recalling (3.20a-b), we write for any $\delta>0$:

$$c(T-T_0)\geqq\frac{c}{1+\delta}(T-T_0)+\frac{C_f}{T}\geqq\frac{c}{1+\delta}(T-T_0)+f(T) \tag{3.24}$$

for T sufficiently large, in fact, $T>T_\delta$

$$T_\delta=\frac{T_0+\sqrt{T_0^2+4C_f(1+\delta)/c\delta}}{2}\downarrow\frac{T_0+\sqrt{T_0^2+4C_f/c}}{2}$$

as $\delta\uparrow\infty$. Thus, recalling (3.19),

$$\begin{aligned}c(T-T_0)\|\{\phi^0,\phi^1\}\|_Z^2&\geqq\left[\frac{c}{1+\delta}(T-T_0)+f(T)\right]\|\{\phi^0,\phi^1\}\|_Z^2\\&\geqq\frac{c}{1+\delta}(T-T_0)\|\{\phi^0,\phi^1\}\|_Z^2+2T\|K_{1T}\|^2_{L^2(\Gamma)}.\end{aligned} \tag{3.25}$$

Thus, (3.23) and (3.25) imply (3.18) as desired, for $T>T_\delta$. The proof of (i) of Theorem 1.2 is complete. □

We show part (ii) of Theorem 1.2 in the next corollary.

COROLLARY 3.3. (i) *For T sufficiently large,* (3.10) (*which characterizes exact controllability of problem* (1.1a-d) *with* $g_1\in H_0^1(0,T;L^2(\Gamma))$ *and* $g_2\equiv 0$ *on the space* $\mathscr{D}(A^{1/4})\times[\mathscr{D}(A^{1/4})]'$ *over* $[0,T]$) *and* (2.11) (*which characterizes exact controllability of problem* (1.1a-d) *with* $g_1\in L^2(\Sigma)$ *and* $g_2\equiv 0$ *on the space* $[\mathscr{D}(A^{1/4})]'\times[\mathscr{D}(A^{3/4})]'$ *over* $[0,T]$) *are equivalent conditions.*

(ii) *By interpolation* [LM1, pp. 64-66], *problem* (1.1a-d) *with controls*

$$g_1\in H_0^{1-\theta}(0,T;L^2(\Gamma)),\quad g_2\equiv 0,\quad 0\leqq\theta\leqq 1,\quad \theta\neq\tfrac{1}{2}, \tag{3.26}$$

$$g_1\in H_{00}^{1/2}(0,T;L^2(\Gamma)),\quad g_2\equiv 0,\quad \theta=\tfrac{1}{2} \tag{3.27}$$

is exactly controllable for T sufficiently large on the space

$$\mathscr{D}(A^{1/4-\theta/2})\times\mathscr{D}(A^{-1/4-\theta/2}), \tag{3.28}$$

where we are using the convention that

$$\mathscr{D}(A^{-\beta}), \qquad \beta\geqq 0 \text{ means } [\mathscr{D}(A^{\beta})]'. \tag{3.29}$$

Proof. (i) Step 4 above shows that (2.11) implies (3.18) for T sufficiently large, and hence (3.10) by Lemma 3.2. The converse follows from

$$2\int_\Sigma\left(\frac{\partial\Delta\phi}{\partial\nu}\right)^2 d\Sigma+2\int_\Sigma|K_{1T}|^2\,d\Sigma\geqq\int_\Sigma\left[-\frac{\partial\Delta\phi}{\partial\nu}+K_{1T}\right]^2 d\Sigma$$

by use of (3.19), (3.20a-b) with T sufficiently large.

(ii) We apply the interpolation Theorem [LM1, Thm. 5.1, p. 27] to the operator $\mathscr{L}_{1T}^{*-1}$ which, by part (i), is bounded between the space in (3.28) and the space in (3.26) at the endpoint values $\theta=0$ and $\theta=1$. Hence $\mathscr{L}_{1T}^{*-1}$ is continuous

$$[\mathscr{D}(A^{1/4})\times\mathscr{D}(A^{-1/4}),\ \mathscr{D}(A^{-1/4})\times\mathscr{D}(A^{-3/4})]_\theta\to[H_0^1(0,T;L^2(\Gamma)),L^2(\Sigma)]_\theta, \tag{3.30}$$

which means that $\mathscr{L}_{1T}$ is onto in the opposite direction. Then [LM1, pp. 64-66] gives (3.26)-(3.27). □

Remark 3.1. With reference to Remark 1.1, we show next that the space of exact controllability $H_0^1(\Omega)\times H^{-1}(\Omega)$ with controls $g_1\in H_0^1(0,T;L^2(\Gamma))$ and $g_2\equiv 0$ as in Theorem 1.2 does *not* coincide with the space of regularity of the solutions of the corresponding problem (1.1a-d). More precisely, we have that the map

$$g_1\to\begin{cases}w(t)-G_1g_1(t),\\ w_t(t)\end{cases} \tag{3.31}$$

for problem (1.1a-d) with $g_2\equiv 0$, $w^0=w^1=0$, is continuous $H_0^1(0,T;L^2(\Gamma))\to C([0,T];H_0^1(\Omega)\times H^{-1}(\Omega))$, while $G_1g_1(t)\in H_0^1(0,T;H^{1/2}(\Omega))$ by elliptic theory and $G_1g_1(T)=0$ so that $\{w(T),w_t(T)\}\in H_0^1(\Omega)\times H^{-1}(\Omega)$.

In fact, if we return to (2.3) we have, after an integration by parts in t, using standard properties of cosine/sine operators:

$$w(t)=[C(t-\tau)G_1g(\tau)]_0^t-\int_0^t C(t-\tau)G_1\dot g_1(\tau)\,d\tau,$$

$$w_t(t)=A\left\{[S(t-\tau)G_1g_1(t)]_0^t-\int_0^t S(t-\tau)G_1\dot g_1(\tau)\,d\tau\right\},\quad\text{i.e.,}$$

$$w(t)-G_1g(t)=-\int_0^t C(t-z)G_1\dot g_1(\tau)\,d\tau, \tag{3.32}$$

$$w_t(t)=A\int_0^t S(t-z)G_1\dot g_1(\tau)\,d\tau. \tag{3.33}$$

Since $\dot g_1\in L^2(\Sigma)$, the (optimal) regularity theory of Theorem 1.0 together with (2.3) gives that (see (1.7))

$$w_t(t)=A\int_0^t S(t-\tau)G_1\dot g_1(\tau)\,d\tau\in C([0,T];[\mathscr{D}(A^{1/4})]'=H^{-1}(\Omega)), \tag{3.34}$$

$$A\int_0^t C(t-\tau)G_1\dot g_1(\tau)\,d\tau\in C([0,T];[\mathscr{D}(A^{3/4})]'), \tag{3.35}$$

and hence (see (1.7))

$$w(t)-G_1g_1(t)=-A^{-1}A\int_0^t C(t-\tau)G_1\dot{g}_1(\tau)\,d\tau\in C([0,T];\mathcal{D}(A^{1/4})=H_0^1(\Omega)) \tag{3.36}$$

as desired.

4. Proof of Theorem 1.3. We parallel and complement the proof of Theorem 1.1.

Step 0. We introduce a second Green map G_2 defined by

$$G_2g=y\leftrightarrow\begin{cases}\Delta^2y=0 & \text{in }\Omega, \quad (4.1a)\\ y|_\Gamma=0, & (4.1b)\\ \partial y/\partial\nu|_\Gamma=g, & (4.1c)\end{cases}$$

which is continuous $L^2(\Gamma)\to L^2(\Omega)$ (indeed, $L^2(\Gamma)\to H^{3/2}(\Omega)$ [LM1, Vol. I, pp. 188-189]). Then we define the operator $\mathcal{L}_{2T}$ by

$$\mathcal{L}_{2T}g_2=\begin{vmatrix} A\displaystyle\int_0^T S(T-t)G_2g_2(t)\,dt,\\ A\displaystyle\int_0^T C(T-t)G_2g_2(t)\,dt\end{vmatrix} \tag{4.2}$$

(in the notation of (2.3)). The solution to problem (1.1a-d) with

$$g_1\in L^2(\Sigma),\qquad g_2\in L^2(0,T;H^{-1}(\Gamma)) \tag{4.3}$$

is likewise given by

$$\begin{vmatrix} w(T;\,t=0;\,w^0=0,\,w^1=0)\\ w_t(T,\,t=0;\,w^0=0,\,w^1=0)\end{vmatrix}=\mathcal{L}_{1T}g_1+\mathcal{L}_{2T}g_2. \tag{4.4}$$

The following lemma can be proved as the corresponding Lemma 2.0 by Green's second theorem, (4.1a-c), and (2.2). Therefore details are omitted.

LEMMA 4.0. *Let G_2^* be the continuous operator $L^2(\Omega)\to L^2(\Gamma)$, which is the adjoint of G_2: $(G_2g,v)_\Omega=(g,G_2^*v)_\Gamma$, $g\in L^2(\Gamma)$, $v\in L^2(Q)$. Then*

$$G_2^*Af=-\Delta f|_\Gamma,\qquad f\in\mathcal{D}(A). \tag{4.5}$$

Step 1. The (regularity) Theorem 1.0 gives a fortiori that the operator

$$\mathcal{L}_T=[\mathcal{L}_{1T},\mathcal{L}_{2T}] \tag{4.6}$$

is continuous $L^2(\Sigma)\times L^2(0,T;H^{-1}(\Gamma))\to X\equiv[\mathcal{D}(A^{1/4})]'\times[\mathcal{D}(A^{3/4})]'$, and thus exact controllability of problem (1.1a-d), (4.3) on the space X over $[0,T]$ is equivalent to the following. There is $C_T>0$ such that

$$\begin{aligned}\left\|\mathcal{L}_T^*\begin{vmatrix}z_1\\ z_2\end{vmatrix}\right\|^2_{L^2(\Sigma)\times L^2(0,T;H^{-1}(\Gamma))}&=\left\|\mathcal{L}_{1T}^*\begin{vmatrix}z_1\\ z_2\end{vmatrix}\right\|^2_{L^2(\Sigma)}+\left\|\mathcal{L}_{2T}^*\begin{vmatrix}z_1\\ z_2\end{vmatrix}\right\|^2_{L^2(0,T;H^{-1}(\Gamma))}\\ &\geqq C_T\|\{z_1,z_2\}\|^2_{[\mathcal{D}(A^{1/4})]'\times[\mathcal{D}(A^{3/4})]'}\end{aligned} \tag{4.7}$$

as it follows by using (4.6). In (4.7), $*$ denotes Hilbert space adjoint. Since $\mathcal{L}_{1T}^*$ is identified by (2.8), (2.9a-d), we proceed to characterize $\mathcal{L}_{2T}^*$. The counterpart of Lemma 2.1 is Lemma 4.1.

LEMMA 4.1. (i) *For $z=\{z_1,z_2\}\in X\equiv[\mathcal{D}(A^{1/4})]'\times[\mathcal{D}(A^{3/4})]'$, we have*

$$\left(\mathcal{L}_{2T}^*\begin{vmatrix}z_1\\ z_2\end{vmatrix}\right)(t)=-\Lambda^2\Delta\phi(t,\phi^0,\phi^1)|_\Sigma\quad \text{on }\Sigma, \tag{4.8}$$

where $\phi(t, \phi^0, \phi^1)$ *solves problem* (2.9a–d), (2.10) *and where*

(4.9) Λ *is isomorphism* $H^s(\Gamma) \to H^{s-1}(\Gamma)$ *and self-adjoint on* $L^2(\Gamma)$

(*first-order tangential operator on* Γ *with smooth coefficients*). *Hence, if* ∇_σ *denotes the tangential gradient, we have*

$$\text{(4.10a)} \qquad \left\| \mathscr{L}_{2T}^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right\|^2_{L^2(0,T;H^{-1}(\Gamma))} = \|\Lambda \Delta \phi(\cdot, \phi^0, \phi^1)\|^2_{L^2(\Sigma)}$$

$$\text{(4.10b)} \qquad = \|\|\nabla_\sigma(\Delta\phi)\|\|^2_{L^2(\Sigma)} + \|\Delta\phi\|^2_{L^2(\Sigma)}.$$

(ii) *For any* $0 < T < \infty$, (4.7) *is equivalent to saying*: *There is* $C'_T > 0$ *such that*

$$\text{(4.11)} \qquad \begin{aligned} &\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_\Sigma |\nabla_\sigma(\Delta\phi)|^2 \, d\Sigma + \int_\Sigma |\Delta\phi|^2 \, d\Sigma \\ &\quad = \int_\Sigma |\nabla(\Delta\phi)|^2 \, d\Sigma + \int_\Sigma |\Delta\phi|^2 \, d\Sigma \geqq C'_T \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4}) \times \mathscr{D}(A^{1/4})}. \end{aligned}$$

Proof. (i) By definition of $\mathscr{L}_{2T}^*$ and Λ we have

$$\text{(4.12)} \qquad \begin{aligned} (\mathscr{L}_{2T} g_2, z)_X &= (g_2, \mathscr{L}_{2T}^* z)_{L^2(0,T;H^{-1}(\Gamma))} \\ &= (\Lambda^{-1} g_2, \Lambda^{-1} \mathscr{L}_{2T}^* z)_{L^2(\Sigma)} \\ &= (g_2, \Lambda^{-2} \mathscr{L}_{2T}^* z)_{L^2(\Sigma)}. \end{aligned}$$

On the other hand, starting from (4.2) we compute as usual:

$$\text{(4.13)} \qquad \left(\mathscr{L}_{2T} g_2, \begin{vmatrix} z_1 \\ z_2 \end{vmatrix}\right)_X = \int_0^T (g_2(t), G_2^*[C(T-t)A^{-1/2} z_2 + S(T-t)A^{1/2} z_1])_\Gamma \, dt,$$

and hence, using (4.12) and (4.13), since $C(\cdot)$ is even and $S(\cdot)$ is odd we obtain

$$\text{(4.14)} \qquad \Lambda^{-2} \mathscr{L}_{2T}^* z = G_2^* A[C(t-T)A^{-3/2} z_2 - S(t-T)A^{-1/2} z_1] = G_2^* A \phi(t, \phi^0, \phi^1)$$

with $\phi(t, \phi^0, \phi^1)$ solution of (2.9a–d), (2.10). Thus, (4.14) leads to (4.8), as desired, by virtue of (4.5). Moreover, since by (4.9) we have

$$\text{(4.15)} \qquad \|f\|_{H^{-1}(\Gamma)} = \|\Lambda^{-1} f\|_{L^2(\Gamma)},$$

then (4.15) applied to (4.8) yields (4.10a). To obtain (4.10b) we first note that by (4.9) we have

$$\text{(4.16)} \qquad \|\Lambda\psi\|^2_{L^2(\Gamma)} = \|\psi\|^2_{H^1(\Gamma)} = \|\|\nabla_\sigma \psi\|\|^2_{L^2(\Gamma)} + \|\psi\|^2_{L^2(\Gamma)},$$

where ∇_σ denotes the tangential gradient. More specifically, at each point of Γ, with ν the unit outward normal and $\tau_1, \cdots, \tau_{n-1}$ on orthogonal systems of unit vectors on the tangent plane, we may write:

$$\text{(4.17a)} \qquad \nabla\psi = (\nabla\psi \cdot \nu)\nu + \sum_{i=1}^{n-1} (\nabla\psi \cdot \tau_i)\tau_i = \frac{\partial\psi}{\partial\nu}\nu + \nabla_\sigma \psi, \qquad \nabla_\sigma \psi = \sum_{i=1}^{n-1} \frac{\partial\psi}{\partial\tau_i} \tau_i,$$

$$\text{(4.17b)} \qquad \|\|\nabla\psi\|\|^2_{L^2(\Gamma)} = \left\|\frac{\partial\psi}{\partial\nu}\right\|^2_{L^2(\Gamma)} + \|\|\nabla_\sigma\psi\|\|^2_{L^2(\Gamma)}.$$

Then using (4.16) with $\psi = \Delta\phi$ in (4.10a) yields (4.10b).

(ii) We use (4.7) along with (2.8) and (4.10b), and (4.17b) with $\psi = \Delta\phi$; finally we use (2.18). This way (4.11) is obtained. □

Step 2. The key of the controllability problem in Theorem 1.3 is the following proposition.

PROPOSITION 4.2. *Inequality* (4.11) *holds true for any* $T>0$ *where, by time reversal in* (2.9a-d), (2.10), *we may take* ϕ *to be the solution of the homogeneous problem* (2.20a)-(2.20d).

Remark 4.1. The inequality opposite to (4.11), i.e.,

$$\int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma + \int_\Sigma |\Delta\phi|^2\, d\Sigma \leqq CT\|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{4.18}$$

is true for all $0<T<\infty$ (transposition on [L2]).

Proof of Proposition 4.2. We return to the fundamental identity (2.24) with $h(x)=x-x_0$, for some $x_0\in R^n$.

Step (i). Setting, as in (2.40), $2C_h=\max_\Gamma |h|$, we have for the LHS of (2.24):

$$\begin{aligned}\text{LHS of (2.24)} &\leqq C_h\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 + |\nabla(\Delta\phi)|^2\, d\Sigma \\ &\quad + \frac{n}{4}\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 + |\Delta\phi|^2\, d\Sigma + C_h\int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma \\ &\leqq \left(3C_h+\frac{n}{4}\right)\int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma + \frac{n}{4}\int_\Sigma |\Delta\phi|^2\, d\Sigma \quad \text{(by (4.17b))} \\ &= c_{n,h}\int_\Sigma |\nabla(\Delta\phi)|^2 + |\Delta\phi|^2\, d\Sigma.\end{aligned} \tag{4.19}$$

Step (ii). As to the RHS of (2.24), we invoke the norm-equivalence (2.32)-(2.33) and write

$$\begin{aligned}\text{RHS of (2.24)} &\geqq c\int_0^T \|A^{1/4}\phi_t\|^2_\Omega + \|A^{3/4}\phi\|^2_\Omega\, dt + \beta_{0,T} \\ &= cT\|\{\phi^0, \phi^1\}\|^2_Z + \beta_{0,T},\end{aligned} \tag{4.20}$$

where in the last step we have used the norm-preserving identity (2.29), and where $Z=\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})$ as in (2.28). Now, however, we estimate β_{0T} in a manner different from Step (vi) in the proof of Theorem 1.1 (see (2.39)).

LEMMA 4.3. *With reference to* (2.25) *we have for any* $\varepsilon>0$,

$$|\beta_{0,T}| \leqq \frac{C_{1h,n}}{\varepsilon}[\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))}] + 2CC_{2n}\varepsilon\|\{\phi^0, \phi^1\}\|^2_Z, \tag{4.21}$$

with Z *as in* (2.28), *where* C *is the constant of norm equivalence in* (2.32)-(2.33) *and*

$$C_{1h,n}=\max\left\{\frac{n}{2}, 2M_h\right\}, \quad C_{2n}=\max\left\{\frac{n}{2}, 1\right\}, \quad 2M_h=\max_{\bar\Omega}|h|.$$

Proof of Lemma 4.3. We estimate each term of (2.25) separately. By the Schwarz inequality and (2.32), (2.33), we obtain

$$\begin{aligned}\left|\left[\int_\Omega \nabla\phi\cdot\nabla\phi_t\, d\Omega\right]_0^T\right| &\leqq \|\|\nabla\phi(T)\|\|_\Omega\|\|\nabla\phi_t(T)\|\|_\Omega + \|\|\nabla\phi^0\|\|_\Omega\|\|\nabla\phi^1\|\|_\Omega \\ &\leqq \frac{1}{2}\left\{\frac{1}{\varepsilon}[\|\|\nabla\phi(T)\|\|^2_\Omega + \|\|\nabla\phi^0\|\|^2_\Omega] + \varepsilon[\|\|\nabla\phi_t(T)\|\|^2_\Omega + \|\|\nabla\phi^1\|\|^2_\Omega]\right\} \\ &\leqq \frac{1}{\varepsilon}\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))} + \frac{\varepsilon}{2}\{\|\|\nabla\phi_t(T)\|\|^2_\Omega + \|\|\nabla\phi^1\|\|^2_\Omega\}.\end{aligned} \tag{4.22}$$

Similarly with $2M_h \equiv \max_{\bar{\Omega}} |h|$:

$$[(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T \leqq 2M_h\{\|\phi_t(T)\|_\Omega |||\nabla(\Delta\phi(T))|||_\Omega + \|\phi^1\|_\Omega |||\nabla(\Delta\phi^0)|||_\Omega\}$$

$$\leqq M_h\left\{\frac{1}{\varepsilon}[\|\phi_t(T)\|_\Omega^2 + \|\phi^1\|_\Omega^2] + \varepsilon[|||\nabla(\Delta\phi(T))|||_\Omega^2 + |||\nabla(\Delta\phi^0)|||_\Omega^2]\right\} \tag{4.23}$$

$$\leqq \frac{2M_h}{\varepsilon}\|\phi_t\|^2_{C([0,T];L^2(\Omega))} + M_h\varepsilon[|||\nabla(\Delta\phi(T))|||_\Omega^2 + |||\nabla(\Delta\phi^0)|||_\Omega^2].$$

Hence, using (2.25), (4.22), and (4.23), we obtain

$$|\beta_{0,T}| \leqq \frac{C_{1h,n}}{\varepsilon}[|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))}]$$

$$+ C_{2n}\varepsilon[|||\nabla\phi_t(T)|||_\Omega^2 + |||\nabla(\Delta\phi(T))|||_\Omega^2 + |||\nabla\phi^1|||_\Omega^2 + |||\nabla(\Delta\phi^0)|||_\Omega^2]$$

(using the norm equivalence (2.32), (2.33))

$$\leqq (C_{1h,n}/\varepsilon)[|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))}]$$

$$+ CC_{2n}\varepsilon[\|A^{1/4}\phi_t(T)\|_\Omega^2 + \|A^{3/4}\phi(T)\|_\Omega^2 + \|A^{1/4}\phi^1\|_\Omega^2 + \|A^{3/4}\phi^0\|_\Omega^2] \tag{4.24}$$

from which the desired conclusion (4.21) follows by using the norm identity (2.29). □

Step (iii). Combining (4.20) with (4.21), we obtain

$$\text{RHS of (2.24)} \geqq (cT - 2CC_{2n}\varepsilon)\|\{\phi^0, \phi^1\}\|_Z^2$$

$$- \frac{C_{1h,n}}{\varepsilon}[|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} - \|\phi_t\|^2_{C([0,T];L^2(\Omega))}. \tag{4.25}$$

Moreover, combining (4.25) with (4.19) we arrive at

$$\frac{C_{1h,n}}{\varepsilon}[|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))}] + C_{n,h}\int_\Sigma \nabla|(\Delta\phi)|^2 + |\Delta\phi|^2\, d\Sigma$$

$$\geqq c\left(T - \frac{2CC_{2n}\varepsilon}{c}\right)\|\{\phi^0, \phi^1\}\|_Z^2. \tag{4.26}$$

To complete the proof of Proposition 4.2, we need the following lemma based on compactness arguments, of the type already used in Littman [L9], who invokes Hormander [H2], in Lions [L3], who uses a remark of P. L. Lions, and in [LT3] in the context of the wave equations. This lemma consists in "absorbing" the lower-order interior terms on the left of (4.26) by the boundary terms on the left of (4.26).

LEMMA 4.4. *Inequality* (4.26) *implies that for any* $T > 0$ *there exists* $C_T > 0$ *such that*

$$|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))} \leqq C_T\int_\Sigma |\nabla(\Delta\phi)|^2 + |\Delta\phi|^2\, d\Sigma \tag{4.27}$$

(*and indeed, we could replace* $\int_\Sigma |\nabla(\Delta\phi)|^2\, d\Sigma$ *with* $\int_\Sigma |\partial\Delta\phi/\partial\nu|^2\, d\Sigma$ *on the right of* (4.27); *see also Lemma* 5.4 *below*).

Proof of Lemma 4.4. The proof is by contradiction. Let there exist a sequence $\{\phi_n(t)\}$ of solutions to problem (2.20a–d) over $[0, T]$:

$$\phi_n'' + \Delta^2\phi_n = 0 \qquad \text{in } Q, \tag{4.28a}$$

$$\phi_n(0,\cdot) = \phi_n^0 \in \mathcal{D}(A^{3/4}), \quad \phi_n'(0,\cdot) = \phi_n^1 \in \mathcal{D}(A^{1/4}) \qquad \text{in } \Omega, \tag{4.28b}$$

$$\phi_n|_\Sigma \equiv 0 \qquad \text{in } \Sigma, \tag{4.28c}$$

$$\left.\frac{\partial\phi_n}{\partial\nu}\right|_\Sigma \equiv 0 \qquad \text{in } \Sigma \tag{4.28d}$$

$(d/dt = ')$, given explicitly by

$$\phi_n(t) = C(t)\phi_n^0 + S(t)\phi_n^1 \in C([0, T]; \mathcal{D}(A^{3/4})), \tag{4.29a}$$

$$\phi_n'(t) = -AS(t)\phi_n^0 + C(t)\phi_n^1 \in C([0, T]; \mathcal{D}(A^{1/4})) \tag{4.29b}$$

such that

$$|||\nabla \phi_n|||_{C([0,T];L^2(\Omega))} \equiv 1, \tag{4.30a}$$

$$\|\phi_n'\|_{C([0,T];L^2(\Omega))} \equiv 1, \tag{4.30b}$$

$$\int_\Sigma |\nabla(\Delta\phi_n)|^2 + |\Delta\phi_n|^2 \, d\Sigma \to 0 \quad \text{as } n \to \infty. \tag{4.30c}$$

By the preceding steps (i)-(iii), each solution $\phi_n(t)$ satisfies (4.26), and thus we have

$$\|\{\phi_n^0, \phi_n^1\}\|_{\mathcal{D}(A^{3/4})\times\mathcal{D}(A^{1/4})} \leqq \text{const} \quad \text{uniformly in } n. \tag{4.31}$$

Hence there is a subsequence, still subindexed by n, such that

$$\phi_n^0 \to \text{some function } \phi^0 \text{ in } \mathcal{D}(A^{3/4}) \quad \text{weakly}, \tag{4.32a}$$

$$\phi_n^1 \to \text{some function } \phi^1 \text{ in } \mathcal{D}(A^{1/4}) \quad \text{weakly}. \tag{4.32b}$$

We then consider the solution to problem (2.20a-d) with initial data found in (4.32a, b):

$$\tilde{\phi}(t) = C(t)\phi^0 + S(t)\phi^1 \in C([0, T]; \mathcal{D}(A^{3/4}) = V), \tag{4.33a}$$

$$\tilde{\phi}'(t) = -AS(t)\phi^0 + C(t)\phi^1 \in C([0, T]; \mathcal{D}(A^{1/4}) = H_0^1(\Omega)). \tag{4.33b}$$

Then (see details, e.g., in [LT3, § 2] in a similar situation corresponding to the wave equation), it follows that

$$\phi_n(t) \to \tilde{\phi}(t) \text{ in } L^\infty(0, T; \mathcal{D}(A^{3/4})) \quad \text{weak star}, \tag{4.34a}$$

$$\phi_n'(t) \to \tilde{\phi}'(t) \text{ in } L^\infty(0, T; \mathcal{D}(A^{1/4})) \quad \text{weak star}. \tag{4.34b}$$

Then (4.34a, b) implies that $\phi_n(t)$ and $\phi_n'(t)$ are uniformly bounded in $L^\infty(0, T; \mathcal{D}(A^{3/4}))$ and $L^\infty(0, T; \mathcal{D}(A^{1/4}))$, respectively. This fact, along with the compactness of $\mathcal{D}(A^{3/4}) \to \mathcal{D}(A^{1/4}) = H_0^1(\Omega)$ (see (2.31)) and of $\mathcal{D}(A^{1/4}) \to L^2(\Omega)$, implies that there is a subsequence, still subindexed by n, such that

$$\phi_n(t) \to \tilde{\phi}(t) \quad \text{strongly in } L^\infty(0, T; H_0^1(\Omega)), \tag{4.35a}$$

$$\phi_n'(t) \to \tilde{\phi}'(t) \quad \text{strongly in } L^\infty(0, T; L^2(\Omega)). \tag{4.35b}$$

A fortiori for (3.30a, b) and (3.35a, b), we obtain

$$\begin{aligned} 1 \equiv |||\nabla \phi_n|||_{C([0,T];L^2(\Omega))} &\to |||\nabla \tilde{\phi}|||_{C([0,T];L^2(\Omega))} = 1, \\ 1 \equiv \|\phi_n'\|_{C([0,T];L^2(\Omega))} &\to \|\tilde{\phi}'\|_{C([0,T];L^2(\Omega))} = 1. \end{aligned} \tag{4.36}$$

Moreover, a fortiori from (4.30c)

$$\left.\frac{\partial(\Delta\tilde{\phi})}{\partial\nu}\right|_\Sigma = 0 \quad \text{and} \quad \Delta\tilde{\phi}|_\Sigma = 0. \tag{4.37}$$

Thus $\tilde{\phi}(t)$ satisfies

$$\tilde{\phi}'' + \Delta^2\tilde{\phi} = 0, \tag{4.38a}$$

$$\tilde{\phi}|_\Sigma \equiv 0, \quad \left.\frac{\partial\tilde{\phi}}{\partial\nu}\right|_\Sigma \equiv 0 \qquad \text{from (4.33a)}, \tag{4.38b}$$

$$\left.\frac{\partial(\Delta\tilde{\phi})}{\partial\nu}\right|_\Sigma \equiv 0, \quad \Delta\tilde{\phi}|_\Sigma \equiv 0 \quad \text{from (4.37)} \tag{4.38c}$$

on $[0, T]$. Then, with $T>0$ arbitrarily small, Holmgren's classical uniqueness theorem of Remark 1.4 implies

$$\tilde{\phi} \equiv 0 \quad \text{in } Q \tag{4.39}$$

and this contradicts (4.36). The proof of Lemma 4.4 is complete. □

Step (iv). We use Lemma 4.4 in (4.26) and obtain for an arbitrary $T>0$ and with ε chosen sufficiently small:

$$C_{\varepsilon,n,T,h} \int_{\Sigma} |\nabla(\Delta\phi)|^2 + |\Delta\phi|^2 \, d\Sigma \geqq c\left[T - \frac{2CC_{2n}\varepsilon}{c}\right] \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{4.40}$$

and (4.40) proves (4.11) for $T>0$ arbitrarily small. The proof of Proposition 4.2 is now complete. □

Then Lemma 4.1, Proposition 4.2, and (4.7) prove Theorem 1.3. □

We conclude this section by providing, with minor extra effort, an extension of Theorem 1.3 to the case when g_2 acts only on a suitable portion of Γ, still with no geometrical conditions on Ω. To this end, given an arbitrary positive number $\gamma>0$ and an arbitrary fixed point $x^0 \in R^n$, we can always decompose Γ into two complementary parts:

$$\Gamma = \Gamma_+(x^0, \gamma) \cup \Gamma_-(x^0, \gamma), \tag{4.41}$$

$$\Gamma_+(x^0, \gamma) = \{x \in \Gamma : (x - x^0) \cdot \nu(x) \geqq \gamma > 0\}, \tag{4.42}$$

$$\Gamma_-(x^0, \gamma) = \{x \in \Gamma : (x - x^0) \cdot \nu(x) < \gamma\}. \tag{4.43}$$

Henceforth we shall drop the explicit dependence on γ and write more simply $\Gamma_+(x^0)$ and $\Gamma_-(x^0)$. In the next corollary we take g_2 to be active only on $\Gamma_-(x^0)$, assumed nonempty.

THEOREM 4.5. *The exact controllability result of Theorem* 1.3 *remains true for an arbitrary* $T>0$, *and still with no geometrical conditions on* Ω, *with respect to the boundary conditions* $w|_{\Sigma} = g_1 \in L^2(\Sigma)$ *as in Theorem* 1.3 *and*

$$\left.\frac{\partial w}{\partial \nu}\right|_{\Sigma} = \begin{cases} 0 \quad \text{on } (0, T)\times\Gamma_+(x^0) = \Sigma_+(x^0), \\ g_2 \in L^2(0, T; [H^1(\Gamma_-(x^0))]') \end{cases} \tag{4.44}$$

instead of $g_2 \in L^2(0, T: H^{-1}(\Gamma))$ *as in Theorem* 1.3, *with* $\Gamma_-(x^0)$ *nonempty.*

Proof. The proof is a minor variation of the proof of Theorem 1.3. Instead of the operator G_2 in (4.1a-c) we now define the operator $\tilde{G}_2$ by $\tilde{G}_2 g = y$, where y solves problem (4.1a, b) and

$$\left.\frac{\partial y}{\partial \nu}\right|_{\Gamma_+(x^0)} = 0, \qquad \left.\frac{\partial y}{\partial \nu}\right|_{\Gamma_-(x^0)} = g \tag{4.45}$$

instead of (4.1c). Accordingly, (4.5) of Lemma 4.0 now becomes

$$\tilde{G}_2^* A f = -\Delta f|_{\Gamma_-(x^0)}, \qquad f \in \mathscr{D}(A). \tag{4.46}$$

Hence, the counterpart of Lemma 4.1 is as follows. We obtain (4.8), this time restricted only on $\Sigma_-(x^0)$, however, and hence

$$\|\mathscr{L}_{2T}^* z\|^2_{L^2(0,T;[H^1(\Gamma_-(x^0))]')} = \|\|\nabla_\sigma(\Delta\phi)\|\|^2_{L^2(\Sigma_-(x^0))} + \|\Delta\phi\|^2_{L^2(\Sigma_-(x^0))} \tag{4.47}$$

[3] See footnote 2 on (1.8) (p. 331).

instead of (4.10b). Thus, for any $T>0$, the new condition, a counterpart of (4.11), characterizes exact controllability in the present case with control g_2 as in (4.44):

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_{\Sigma_-(x^0)} (|\nabla_\sigma(\Delta\phi)|^2 + |\Delta\phi|^2)\, d\Sigma_-(x^0) \geqq C_T' \|\{\phi^0, \phi^1\}\|_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}, \tag{4.48}$$

where ϕ solves problem (2.9a–d), (2.10) as before. To prove (4.48), we need only to refine slightly the proof of Theorem 1.3 at the level of analyzing the LHS of identity (2.24). We decompose $\nabla(\Delta\phi)$ into its normal and tangential component as in (4.17a, b) with $\psi = \Delta\phi$, so that we obtain

$$\begin{aligned}\text{LHS of (2.24)} = {} & \frac{1}{2}\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 h\cdot\nu\, d\Sigma + \frac{n}{2}\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi\, d\Sigma \\ & + \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla_\sigma(\Delta\phi)\, d\Sigma - \frac{1}{2}\int_\Sigma |\nabla_\sigma(\Delta\phi)|^2 h\cdot\nu\, d\Sigma,\end{aligned} \tag{4.49}$$

where $h(x) = (x - x^0)$.

Next, setting $2M_h^+ = \max|h|$ over $\Gamma_+(x^0)$ and recalling (4.42), we obtain

$$\begin{aligned}&\int_{\Sigma_+(x^0)} \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla_\sigma(\Delta\phi)\, d\Sigma_+(x^0) - \frac{1}{2}\int_{\Sigma_+(x^0)} |\nabla_\sigma(\Delta\phi)|^2 h\cdot\nu\, d\Sigma_+(x^0) \\ &\leqq \frac{(M_h^+)^2}{\varepsilon}\int_{\Sigma_+(x_0)} \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma_+(x^0) + \left(\varepsilon - \frac{\gamma}{2}\right)\int_{\Sigma_+(x^0)} |\nabla_\sigma(\Delta\phi)|^2\, d\Sigma_+(x^0) \\ &\leqq \frac{(M_h^+)^2}{\varepsilon}\int_{\Sigma_+(x^0)} \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma_+(x^0)\end{aligned} \tag{4.50}$$

after selecting $0<\varepsilon<\gamma/2$. Then, by (4.49) and (4.50), if we recall (2.41) we obtain

$$\begin{aligned}&C_{hn\varepsilon}\left\{\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_{\Sigma_-(x^0)} (|\nabla_\sigma(\Delta\phi)|^2 + |\Delta\phi|^2)\, d\Sigma_-(x^0)\right\} \\ &\quad + n\varepsilon_1 c_1 \int_0^T \|A^{3/4}\phi\|_\Omega^2\, dt \geqq \text{LHS of (2.24)},\end{aligned} \tag{4.51}$$

which is the counterpart of (2.45), or of (4.19). Combining (4.51) with (4.25) for the right-hand side of (2.24), yields for any $T>0$ and for any $\varepsilon_1>0$, $\varepsilon>0$

$$\begin{aligned}&C_{1h\varepsilon n}\{|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|_{C([0,T];L^2(\Omega))}\} \\ &\quad + C_{2hn}\left\{\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_{\Sigma_-(x^0)} (|\nabla_\sigma(\Delta\phi)|^2 + |\Delta\phi|^2)\, d\Sigma_-(x^0)\right\} \\ &\geqq [(c - n\varepsilon_1 c_1)T - 2cc_{2n}\varepsilon]\|\{\phi^0, \phi^1\}\|_Z^2,\end{aligned} \tag{4.52}$$

counterpart of (4.26). Next, the counterpart of Lemma 4.4 is that inequality (4.52) implies that for all $T>0$ there exists C_T such that

$$\begin{aligned}&|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))} \\ &\quad \leqq C_T\left\{\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_{\Sigma_-(x^0)} |\Delta\phi|^2\, d\Sigma_-(x^0)\right\}\end{aligned} \tag{4.53}$$

instead of (4.27). Indeed, the contradiction argument of Lemma 4.4 still works, since the corresponding uniqueness property—that $\tilde{\phi}$ in (4.33a, b) satisfies (4.38a, b) as well as

$$\frac{\partial(\Delta\tilde{\phi})}{\partial\nu}\bigg|_{\Sigma} \equiv 0, \qquad \Delta\tilde{\phi}|_{\Sigma_-(x^0)} \equiv 0 \tag{4.54}$$

on $[0, T]$, T arbitrary >0 (instead of (4.38c))—still implies that $\phi \equiv 0$ in Q, by the Holmgren Uniqueness Theorem as in [H2, Thm. 5.33, p. 129] (four homogeneous boundary conditions on a nonempty portion of the boundary are enough). Hence (4.52) and (4.54) imply

$$\begin{aligned} C'_{nh\varepsilon}\left\{\int_{\Sigma}\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \int_{\Sigma_-(x^0)} (|\nabla_\sigma(\Delta\phi)|^2 + |\Delta\phi|^2)\, d\Sigma_-(x^0)\right\} \\ \geqq [(C - n\varepsilon c_1)T - 2CC_{2n}\varepsilon]\|\{\phi^0, \phi^1\}\|_Z^2 \end{aligned} \tag{4.55}$$

for any $T>0$, and any $\varepsilon_1>0$, $\varepsilon>0$ (counterpart of (4.40)). Thus (4.48) is proved. The proof of Theorem 4.5 is complete. □

5. Theorem 1.4. Improvement of Theorem 1.1 to T arbitrarily small. The proof is divided into two steps, which will be treated separately in the following two subsections.

5.1. A preliminary result: exact controllability on the space $X \equiv [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$ with $g_1 \in L^2(\Sigma)$, $g_2 \in H^k(\Sigma)$, k an arbitrary nonnegative integer in arbitrarily short time and with geometrical conditions. A first preliminary result is the following proposition.

PROPOSITION 5.1. *Assume condition* (1.8) *on the domain* Ω. *Then for all* $T>0$, *given* $\{w^0, w^1\} \in X$ *there exist control functions* $g_1 \in L^2(\Sigma)$, $g_2 \in H^k(\Sigma)$, k *a preassigned fixed nonnegative integer, such that the corresponding solution of problem* (1.1) *satisfies*

$$w(T) = w_t(T) = 0, \qquad \{w(t), w_t(t)\} \in C([0, T]; X)$$

Proof of Proposition 5.1.

Step 1. We now return to the operator $\mathscr{L}_T$ defined by (4.6), (4.4), (2.3), and (4.2), and require that

$$\mathscr{L}_T : L^2(\Sigma) \times H^k(\Sigma) \quad \text{onto } X. \tag{5.0}$$

Equivalently, there exists C_T such that

$$\begin{aligned} \left\| \mathscr{L}_T^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right\|^2_{L^2(\Sigma)\times H^k(\Sigma)} &= \left\| \mathscr{L}_{1T}^* \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right\|^2_{L^2(\Sigma)} + \left\| \mathscr{L}_{2T}^{\#} \begin{vmatrix} z_1 \\ z_2 \end{vmatrix} \right\|^2_{H^k(\Sigma)} \\ &\geqq C_T \|\{z_1, z_2\}\|^2_{[\mathscr{D}(A^{1/4})]'\times[\mathscr{D}(A^{3/4})]'}, \end{aligned} \tag{5.1}$$

a counterpart of (4.7), where $\mathscr{L}_{1T}^*$ is identified by (2.8), (2.9a–d) and where $\mathscr{L}_{2T}^{\#}$ is defined by

$$\begin{aligned} (\mathscr{L}_{2T} g_2, z)_X &= (g_2, \mathscr{L}_{2T}^{\#} z)_{H^k(\Sigma)} = (\Lambda_k g_2, \Lambda_k \mathscr{L}_{2T}^{\#} z)_{L^2(\Sigma)} \\ &= (g_2, \Lambda_k^2 \mathscr{L}_{2T}^{\#} z)_{L^2(\Sigma)}, \end{aligned} \tag{5.2}$$

with

$$\Lambda_k : \text{isomorphism } H^k(\Sigma) \quad \text{onto } L^2(\Sigma) \text{ and self-adjoint on } L^2(\Sigma). \tag{5.3}$$

The counterpart of Lemma 4.1 is Lemma 5.2.

LEMMA 5.2. (i) *For* $z=[z_1, z_2]\in X\equiv[\mathscr{D}(A^{1/4})]'\times[\mathscr{D}(A^{3/4})]'$ *we have with reference to* (5.3)

$$(\mathscr{L}^{\#}_{2T}z)(t)=-\Lambda_k^{-2}\Delta\phi(t,\phi^0,\phi^1), \tag{5.4}$$

where $\phi(t,\phi^0,\phi^1)$ *solves problem* (2.9)-(2.10). *Moreover,*

$$\|\mathscr{L}^{\#}_{2T}z\|_{H^k(\Sigma)}=\|\Delta\phi(t,\phi^0,\phi^1)\|_{H^{-k}(\Sigma)}. \tag{5.5}$$

(ii) *For any* $0<T<\infty$, *inequality* (5.1), *which is equivalent to exact controllability in the present case, is in turn equivalent to saying: There is* $C'_T>0$ *such that*

$$\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\|\Delta\phi\|^2_{H^{-k}(\Sigma)}\geqq C'_T\|\{\phi^0,\phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}. \tag{5.6}$$

Proof. (i) Comparison between (5.2) and (4.13) now yields, by (4.5),

$$\Lambda_k^2\mathscr{L}^{\#}_{2T}z=G_2^*A[C(t-T)A^{-3/2}z_2-S(t-T)A^{-1/2}z_1]=-\Delta\phi(t,\phi^0,\phi^1), \tag{5.7}$$

a counterpart of (4.14). Thus, by (5.3) and (5.7),

$$\begin{aligned}\|\mathscr{L}^{\#}_{2T}z\|_{H^k(\Sigma)}&=\|\Lambda_k\mathscr{L}^{\#}_{2T}z\|_{L^2(\Sigma)}=\|\Lambda_k^{-1}\Delta\phi(t,\phi^0,\phi^1)\|_{L^2(\Sigma)}\\&=\|\Delta\phi(t,\phi^0,\phi^1)\|_{H^{-k}(\Sigma)}\end{aligned}$$

as desired.

(ii) Inequality (5.6) follows from (5.1) via (2.8), (2.9a-d), and (5.5). □

Step 2. The key of the present controllability problem is the following proposition, which is the counterpart of Proposition 4.2.

PROPOSITION 5.3. *Under assumption* (1.8), *inequality* (5.6) *holds true for any* $T>0$ *where, by time reversal in* (2.9a-d), (2.10), *we may take* ϕ *to be the solution of the homogeneous problem* (2.20a-d).

Proof of Proposition 5.3. We return to the fundamental identity (2.24) with $h(x)=x-x_0$, for some $x_0\in R^n$. As to the LHS of (2.24), we invoke Step (vii) in § 2, thus obtaining (2.45) under assumption (1.8). Hence the top of (2.47) holds true; we rewrite it here for convenience:

$$C_{1,nh\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+Tn\varepsilon C\|\{\phi^0,\phi^1\}\|^2_Z\geqq\text{LHS of (2.24)}. \tag{5.8}$$

As to the RHS of identity (2.24), we invoke instead Lemma 4.3, and hence (4.25), which we again rewrite here for convenience:

$$\begin{aligned}\text{RHS of (2.24)}\geqq{}&(cT-2CC_{2n}\varepsilon)\|\{\phi^0,\phi^1\}\|^2_Z\\&-\frac{-C_{1h,n}}{\varepsilon}[\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))}-\|\phi_t\|^2_{C([0,T];L^2(\Omega))}].\end{aligned} \tag{5.9}$$

Thus, combining (5.8) with (5.9), we obtain

$$\begin{aligned}&C_{1,nh\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\frac{C_{1hn}}{\varepsilon}[\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))}+\|\phi_t\|^2_{C([0,T];L^2(\Omega))}]\\&\quad\geqq[(c-n\varepsilon C)T-2CC_{2n}\varepsilon]\|\{\phi^0,\phi^1\}\|^2_Z.\end{aligned} \tag{5.10}$$

Then to complete the proof and obtain (5.6) we need the following lemma, which is a counterpart of Lemma 4.4.

LEMMA 5.4. *Inequality* (5.10) *implies that for any* $T>0$ *there exists* $C_T>0$ *such that*

$$\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))}+\|\phi_t\|^2_{C([0,T];L^2(\Omega))}\leqq C_T\left\{\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\|\Delta\phi\|^2_{H^{-k}(\Sigma)}\right\}. \tag{5.11}$$

Proof of Lemma 5.4. The proof is similar to that of Lemma 4.4 to which it reduces for $k=0$. By contradiction, let there exist a sequence $\{\phi_n(t)\}$ of solutions to problem (2.20a-d) = (4.28) over $[0, T]$, given by (4.29) such that

$$|||\nabla \phi_n|||_{C([0,T];L^2(\Omega))} \equiv 1, \tag{5.12a}$$

$$\|\phi_n'\|_{C([0,T];L^2(\Omega))} \equiv 1, \tag{5.12b}$$

$$\int_\Sigma \left(\frac{\partial(\Delta\phi_n)}{\partial\nu}\right)^2 d\Sigma + \|\Delta\phi_n\|^2_{H^{-k}(\Sigma)} \to 0, \qquad n\to\infty. \tag{5.12c}$$

Then, since each $\phi_n(t)$ satisfies inequality (5.10), we obtain (4.31); hence (4.34a, b), (4.35a, b), and (4.36). Moreover, with $\tilde{\phi}$ given by (4.33a, b) we again obtain (4.37) (this time by (5.12c)), and thus (4.38). Hence (4.39) follows and contradicts (4.36). Inequality (5.11) is thus proved. □

5.2. Completion of the proof of Theorem 1.4. We return to inequality (5.10), which is valid for all $T>0$. From here we see that, to obtain (2.11) for any $T>0$ (and hence the desired exact controllability claimed by Theorem 1.4), we need the following improvement of Lemma 5.4.

LEMMA 5.5. (i) *Inequality* (5.10) *implies that for any* $T>0$ *there is* $C_T>0$ *such that*

$$|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))} \leqq C_T \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma. \tag{5.13}$$

(ii) *Thus, under assumption* (1.8) *on* Ω, *for any* $T>0$ *there exists* $C_T>0$ *such that the following inequality holds true for the solution of problem* (2.9a, b), (2.10):

$$\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma \geqq C_T \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})} \tag{5.14}$$

(*improvement from T sufficiently large to T arbitrarily small over* (2.19) *of Lemma* 2.2).

Proof. From the argument of Lemma 5.4, it is clear that (5.13) is indeed achieved, provided the following uniqueness property holds true:

$$\begin{aligned} &\phi_{tt} + \Delta^2\phi \equiv 0, \qquad \phi|_\Sigma \equiv \partial\phi/\partial\nu|_\Sigma \equiv \partial(\Delta\phi)/\partial\nu|_\Sigma \equiv 0 \quad \text{on } [0, T], \\ &\qquad\qquad T>0 \text{ arbitrarily fixed implies } \phi\equiv 0 \quad \text{in } Q, \end{aligned} \tag{5.15}$$

which is needed to obtain the required contradiction. To establish (5.15), we use a recent argument of [Z1], adapted to the present situation, which rests on an idea from [BLR1]. The proof of (5.15) will hinge on the following three results, which have been already obtained.

Result 1. The same uniqueness property holds true as in (5.15), except that T is now greater than some finite $T_0>0$.

As we have seen, this result follows a fortiori from Lemma 2.2, (2.19), under the geometrical condition (1.8) for Ω, with T_0 as in (2.51).

Result 2. Inequality (5.6) is valid for any $T>0$ under assumption (1.8), as guaranteed by Proposition 5.3. (Note that (5.6) was obtained via Lemma 5.4, which we now seek to improve to the form expressed by (5.13).)

Result 3. For any $T>0$, we have

$$\left\|\frac{\partial(\Delta\phi)}{\partial\nu}\right\|^2_{L^2(\Sigma)} \leqq C_T \|\{\phi^0, \phi^1\}\|^2_{\mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4})}. \tag{5.16}$$

Inequality (5.16) follows by transposition on recent results in [L2].

With these three results at hand, and following an idea in [BLR1], we are now in a position to prove (5.15), and hence (5.13). We introduce the space

$$\mathscr{F} \equiv L^\infty(0, T; \mathscr{D}(A^{3/4})) \cap W^{1,\infty}(0, T; \mathscr{D}(A^{1/4})), \tag{5.17}$$

which contains the solutions $\{\phi(t)\}$ of the homogeneous problem (2.9) with data $\{\phi^0, \phi^1\} \in \mathcal{D}(A^{3/4}) \times \mathcal{D}(A^{1/4})$. Also, let

(5.18) $\mathcal{Y} \equiv$ space of all solutions in $\mathcal{F}$ of problem (2.9) that, moreover, satisfy the additional boundary condition $\partial(\Delta\phi)/\partial\nu|_\Sigma \equiv 0$.

The key point is to show that $\mathcal{Y}$ is finite-dimensional. (This is the idea discussed in [BLR1].) To this end, with $\mathcal{Y}$ closed in $\mathcal{F}$ by virtue of (5.16) (Result 3), we seek to establish that

$$B_{\mathcal{F}} \cap \mathcal{Y} \quad \text{is compact,} \tag{5.19}$$

where $B_{\mathcal{F}}$ is the closed unit ball in $\mathcal{F}$ centered at the origin.

Let $\phi \in B_{\mathcal{F}} \cap \mathcal{Y}$. Then we plainly have that ϕ_t satisfies

$$(\phi_t)_{tt} + \Delta^2(\phi_t) \equiv 0 \qquad \text{in } Q, \tag{5.20a}$$

$$\phi_t|_\Sigma \equiv \frac{\partial \phi_t}{\partial \nu}\bigg|_\Sigma \equiv \frac{\partial(\Delta\phi_t)}{\partial \nu} \equiv 0 \quad \text{in } \Sigma. \tag{5.20b}$$

In addition, interior regularity and trace theory imply that $\Delta\phi|_\Sigma \in H^{1/2}(\Sigma)$ and a fortiori

$$\Delta\phi_t \in H^{-1}(0, T; L^2(\Gamma)). \tag{5.21}$$

Next, we use Result 2, i.e., (5.6) of Proposition 5.3 as applied to ϕ_t (which is also a solution of problem (2.9a, b)). We obtain for T arbitrarily short as in (5.15):

$$\begin{aligned} \|\{(\phi_t)^0, (\phi_t)^1\}\|^2_{\mathcal{D}(A^{3/4})\times\mathcal{D}(A^{1/4})} &\leqq C'_T\left\{\int_\Sigma \left(\frac{\partial \Delta\phi_t}{\partial \nu}\right)^2 d\Sigma + \|\Delta\phi_t\|^2_{H^{-k}(\Sigma)}\right\} \\ &\leqq C'_T\|\Delta\phi_t\|_{H^{-1}(0,T;L^2(\Gamma))} \\ &\leqq C_T\|\{\phi^0, \phi^1\}\|_{\mathcal{D}(A^{3/4})\times\mathcal{D}(A^{1/4})}, \end{aligned} \tag{5.22}$$

where in the last step we have used the last boundary condition in (5.20b) as well as (5.21). Thus $\{(\phi_t)^0, (\phi_t)^1\} \in \mathcal{D}(A^{3/4}) \times \mathcal{D}(A^{1/4})$. From (5.22) we deduce (recalling (1.7)) that

$$\phi_t(t) = C(t)(\phi_t)^0 + S(t)(\phi_t)^1 \in C([0, T]; \mathcal{D}(A^{3/4}) \equiv V), \tag{5.23a}$$

$$\phi_{tt}(t) = -AS(t)(\phi_t)^0 + C(t)(\phi_t)^1 \in C([0, T]; \mathcal{D}(A^{1/4}) \equiv H_0^1(\Omega)), \tag{5.23b}$$

i.e.,

$$\phi_t \in \mathcal{F}. \tag{5.24}$$

From the equation satisfied by ϕ, we deduce via (5.23b)

$$\phi_{tt} = -\Delta^2\phi \in C([0, T]; \mathcal{D}(A^{1/4}) \equiv H_0^1(\Omega)). \tag{5.25}$$

Hence,

$$\phi \in C([0, T]; H^5(\Omega)). \tag{5.26}$$

Then (5.26) and (5.23a) prove the following: If $\phi \in B_{\mathcal{F}} \cap \mathcal{Y}$, then

$$\{\phi(t), \phi_t(t)\} \in C([0, T]; H^5(\Omega)) \times C([0, T]; V) \subset C([0, T]; V) \times C([0, T]; H_0^1(\Omega)),$$

where the containment has compact injection. Thus, $\mathcal{Y}$ is finite-dimensional. Then, as in [BLR1], the elements of $\mathcal{Y}$ are solutions of an equivalent finite-dimensional ordinary differential equation with constant coefficients. Since such solutions vanish for $T > T_0 > 0$ by Result 1, they also vanish for $T > 0$ arbitrarily small. Thus property (5.15) follows. □

6. Extension of the proof of § 2 to general vector fields. In this section, we provide the additional arguments—over those of § 2—required to prove Theorem 1.5 in the generality that involves a vector field satisfying conditions (1.21), (1.23) only, not just a radial vector field as in Theorem 1.1 (or a linear vector field as in Remark 1.2).

6.1. Extension of Lemma 2.2. The key point that remains to be established in Proposition 6.1.

PROPOSITION 6.1. *The conclusions of Lemma* 2.2 *with* $T > T_u$ *defined by* (1.24) *hold true for a domain* Ω *satisfying the vector field conditions* (1.21), (1.23) *as well as the uniqueness condition* (1.24).

Proof. Step (i). We return to (A8) and (A10) in Appendix A, where by imposing the boundary conditions (2.20c, d) we obtain the identity:

$$\begin{aligned}
&\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla(\Delta\phi)\,d\Sigma - \frac{1}{2}\int_\Sigma |\nabla(\Delta\phi)|^2 h\cdot\nu\,d\Sigma \\
&\quad = \int_Q H\nabla(\Delta\phi)\cdot\nabla(\Delta\phi)\,dQ + \int_Q H\nabla\phi_t\cdot\nabla\phi_t\,dQ \\
&\quad\quad + \frac{1}{2}\int_Q \{|\nabla\phi_t|^2 - |\nabla(\Delta\phi)|^2\}\,\mathrm{div}\,h\,dQ \\
&\quad\quad + \int_Q \phi_t\nabla(\mathrm{div}\,h)\cdot\nabla\phi_t\,dQ - [(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T,
\end{aligned} \tag{6.1}$$

counterpart of identity (2.22).

Step (ii). We use identity (B4) in Appendix B. Inserting (B4) in (6.1) yields the identity

$$\begin{aligned}
&\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla(\Delta\phi)\,d\Sigma + \frac{1}{2}\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi\,\mathrm{div}\,h\,d\Sigma - \frac{1}{2}\int_\Sigma |\nabla(\Delta\phi)|^2 h\cdot\nu\,d\Sigma \\
&\quad = \int_Q H\nabla(\Delta\phi)\cdot\nabla(\Delta\phi)\,dQ + \int_Q H\nabla\phi_t\cdot\nabla\phi_t\,dQ \\
&\quad\quad + \frac{1}{2}\int_Q \Delta\phi\,\nabla(\mathrm{div}\,h)\cdot\nabla(\Delta\phi)\,dQ + \frac{1}{2}\int_Q \phi_t\nabla(\mathrm{div}\,h)\cdot\nabla\phi_t\,dQ + \beta_{0,T,h},
\end{aligned} \tag{6.2}$$

where

$$\beta_{0,T,h} = \frac{1}{2}\left[\int_\Omega \nabla\phi\cdot\nabla(\phi_t\,\mathrm{div}\,h)\,d\Omega\right]_0^T - [(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T, \tag{6.3}$$

counterpart of (2.24)-(2.25). Equation (6.2) shows the presence of two new terms over (2.24). We shall now deal with them.

Step (iii). Let

$$4G_h \equiv \max_{\bar\Omega} |\nabla(\mathrm{div}\,h)|. \tag{6.4}$$

Then, for any $\varepsilon > 0$,

$$\begin{aligned}
\frac{1}{2}\int_Q \Delta\phi\,\nabla(\mathrm{div}\,h)\cdot\nabla(\Delta\phi)\,dQ &\geqq -2G_h\int_Q |\Delta\phi||\nabla(\Delta\phi)|\,dQ \\
&\geqq -\varepsilon G_h\int_Q |\nabla(\Delta\phi)|^2\,dQ - \frac{G_h}{\varepsilon}\int_Q |\Delta\phi|^2\,dQ.
\end{aligned} \tag{6.5}$$

Similarly,

$$\begin{aligned}\frac{1}{2}\int_Q \phi_t \nabla(\operatorname{div} h)\cdot\nabla\phi_t\,dQ &\geqq -2G_h\int_Q |\phi_t||\nabla\phi_t|\,dQ \\ &\geqq -\varepsilon G_h\int_Q |\nabla\phi_t|^2\,dQ - \frac{G_h}{\varepsilon}\int_Q \phi_t^2\,dQ.\end{aligned} \tag{6.6}$$

Therefore, if we return to (6.2), use (6.5)-(6.6), and recall assumption (1.23), we see that the RHS of (6.2) satisfies

$$\begin{aligned}\text{RHS of (6.2)} &\geqq (\rho-\varepsilon G_h)\int_Q |\nabla(\Delta\phi)|^2+|\nabla\phi_t|^2\,dQ \\ &\quad -\frac{G_h}{\varepsilon}\int_Q |\Delta\phi|^2+\phi_t^2\,dQ+\beta_{0,T,h} \\ &\geqq (\rho-\varepsilon G_h)C_1\int_0^T \|A^{3/4}\phi\|_\Omega^2+\|A^{1/4}\phi_t\|^2\,dt \\ &\quad -\frac{G_h}{\varepsilon}\int_Q |\Delta\phi|^2+\phi_t^2\,dQ+\beta_{0,T,h},\end{aligned} \tag{6.7}$$

where we have used the norm-equivalence (2.32)-(2.33) of Lemma 2.3(ii), (iii) with constant C_1 of equivalence.

Next, the very same proof as in Lemma 4.3, applied to (6.3) rather than to (2.25), produces for any $\varepsilon'>0$

$$|\beta_{0,T,h}| \leqq \frac{C_{1h}}{\varepsilon'}[|||\nabla\phi|||^2_{C([0,T];L^2(\Omega))}+\|\phi_t\|^2_{C([0,T];L^2(\Omega))}+\varepsilon' C_{2h}\|\{\phi^0,\phi^1\}\|_Z^2, \tag{6.8}$$

$$Z \equiv \mathscr{D}(A^{3/4})\times\mathscr{D}(A^{1/4}) \quad \text{as in (2.28)} \tag{6.9}$$

(counterpart of (4.21)). Combining now (6.8) with (6.7) and using the time invariance (2.29), we finally obtain via (6.9):

$$\begin{aligned}\text{RHS of (6.2)} &\geqq [(\rho-\varepsilon G_h)C_1 T-\varepsilon' C_{2h}]\|\{\phi^0,\phi^1\}\|_Z^2-\frac{G_h}{\varepsilon}\int_Q |\Delta\phi|^2+\phi_t^2\,dQ \\ &\quad -\frac{C_{1h}}{\varepsilon'}[|||\nabla\phi|||^2_{C([0,T];\,L^2(\Omega))}+\|\phi_t\|^2_{C([0,T];\,L^2(\Omega))}],\end{aligned} \tag{6.10}$$

counterpart of (2.46) or (4.25). Equation (6.10) shows the presence of new terms due to $G_h\neq 0$ over (2.46) or (4.25). How to deal with them represents the main additional difficulty over the proof of § 2.

Step (iv). For the LHS of (6.2) we recall (2.40)

$$\left|\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\,h\cdot\nabla(\Delta\phi)\,d\Sigma\right| \leqq \frac{C_h}{\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\varepsilon C_h\int_\Sigma |\nabla(\Delta\phi)|^2\,d\Sigma \tag{6.11}$$

and likewise

$$\begin{aligned}\left|\frac{1}{2}\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi\,\operatorname{div} h\,d\Sigma\right| &\leqq 2D_{h,b}\int_\Sigma\left|\frac{\partial(\Delta\phi)}{\partial\nu}\right||\Delta\phi|\,d\Sigma \\ &\leqq \frac{D_{h,b}}{\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\varepsilon D_{h,b}\int_\Sigma |\Delta\phi|^2\,d\Sigma\end{aligned} \tag{6.12}$$

with $D_{h,b} = \max_\Gamma |\operatorname{div} h|$, $b =$ boundary. Using (6.11)-(6.12) on the LHS of (6.2), we obtain

$$\left(\frac{C_h + D_{h,b}}{\varepsilon}\right) \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \varepsilon D_{h,b} \int_\Sigma |\Delta\phi|^2 \, d\Sigma \geqq \text{LHS of (6.2)}, \tag{6.13}$$

by selecting $\varepsilon > 0$ small enough that

$$\varepsilon C_h \int_\Sigma |\nabla(\Delta\phi)|^2 \, d\Sigma - \frac{1}{2} \int_\Sigma |\nabla(\Delta\phi)|^2 h \cdot \nu \, d\Sigma \leqq \left(\varepsilon C_h - \frac{\gamma}{2}\right) \int_\Sigma |\nabla(\Delta\phi)|^2 \, d\Sigma \leqq 0 \tag{6.14}$$

in view of assumption (1.21).

We now combine (6.13) with (6.10) and obtain

$$\begin{aligned} &\left(\frac{C_h + D_{h,b}}{\varepsilon}\right) \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \varepsilon D_{h,b} \int_\Sigma |\Delta\phi|^2 \, d\Sigma + \frac{G_h}{\varepsilon} \int_Q |\Delta\phi|^2 + \phi_t^2 \, dQ \\ &\qquad + \frac{C_{1h}}{\varepsilon'} [\|\|\nabla\phi\|\|^2_{C([0,T];\, L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];\, L^2(\Omega))}] \qquad (6.15) \\ &\geqq [(\rho C_1 - \varepsilon G_h C_1) T - \varepsilon' C_{2h}] \|\{\phi^0, \phi^1\}\|^2_Z, \end{aligned}$$

counterpart of (2.48).

Step (v). Since $\phi(t) \in \mathcal{D}(A^{3/4}) \subset \mathcal{D}(A^{1/4})$ (see Remark 2.2) and $\phi(t)$ satisfies the boundary conditions (2.20c, d), Green's second theorem yields

$$\|A^{1/2}\phi\|^2_\Omega = (A\phi, \phi)_\Omega = (\Delta(\Delta\phi), \phi)_\Omega = (\Delta\phi, \Delta\phi)_\Omega = \int_\Omega |\Delta\phi|^2 \, d\Omega. \tag{6.16}$$

Thus, recalling the norm-equivalence (2.33), and using (6.16), we obtain

$$\int_\Omega |\nabla\phi|^2 \, d\Omega \leqq c\|A^{-1/4}\phi\|^2_\Omega \leqq c\|A^{-1/4}\|^2 \|A^{1/2}\phi\|^2_\Omega = C \int_\Omega |\Delta\phi|^2 \, d\Omega = C\|\Delta\phi\|^2_\Omega \tag{6.17}$$

so that

$$\|\|\nabla\phi\|\|^2_{C([0,T];L^2(\Omega))} \leqq C\|\Delta\phi\|^2_{C([0,T];L^2(\Omega))}. \tag{6.18}$$

Moreover,

$$\int_Q |\Delta\phi|^2 + \phi_t^2 \, dQ \leqq T\{\|\Delta\phi\|^2_{C([0,T];L^2(\Omega))} + \|\phi_t\|^2_{C([0,T];L^2(\Omega))}\}. \tag{6.19}$$

Using (6.18)-(6.19) in (6.15) yields

$$\begin{aligned} &\left(\frac{C_h + D_{h,b}}{\varepsilon}\right) \int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma + \varepsilon D_{h,b} \int_\Sigma |\Delta\phi|^2 \, d\Sigma \\ &\qquad + \left(\frac{G_h T}{\varepsilon} + \frac{CC_{1h}}{\varepsilon'}\right) \|\Delta\phi\|^2_{C([0,T];L^2(\Omega))} + \left(\frac{G_h T}{\varepsilon} + \frac{C_{1h}}{\varepsilon'}\right) \|\phi_t\|^2_{C([0,T];L^2(\Omega))} \qquad (6.20) \\ &\geqq [(\rho C_1 - \varepsilon G_h C_1) T - \varepsilon' C_{2h}] \|\{\phi^0, \phi^1\}\|^2_Z. \end{aligned}$$

Thus we are in a situation similar (but not identical) to the one we encountered in (4.26) in the proof of Proposition 4.2. We then need a lemma, the counterpart of Lemma 4.4, that will allow us to "absorb" the interior terms on Q on the left of (6.20)

by the boundary terms on Σ on the left of (6.20). This is indeed provided by the following step.

Step (vi). We proceed to Lemma 6.2.

LEMMA 6.2. *Inequality* (6.20) *for problem* (2.20a-d) *implies that, for any* $T>0$, *there exists a constant* $C_T>0$ *such that*

$$\|\Delta\phi\|^2_{C([0,T];L^2(\Omega))}+\|\phi_t\|^2_{C([0,T];L^2(\Omega))}\leqq C_T\left\{\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\int_\Sigma|\Delta\phi|^2\,d\Sigma\right\}. \tag{6.21}$$

Proof of Lemma 6.2. The proof is similar to that of Lemma 4.4. Thus only the relevant differences from the latter proof will be noted here. Suppose that there exists a sequence $\{\phi_n(t)\}$ of solution to problem (2.20) over $[0, T]$, as in (4.28a-d), (4.29a, b) such that

$$\|\Delta\phi_n\|_{C([0,T];L^2(\Omega))}\equiv 1, \tag{6.22a}$$

$$\|\phi_n'\|_{C([0,T];L^2(\Omega))}\equiv 1, \tag{6.22b}$$

$$\int_\Sigma\left(\frac{\partial(\Delta\phi_n)}{\partial\nu}\right)^2+|\Delta\phi_n|^2\,d\Sigma\to 0 \quad \text{as } n\to\infty. \tag{6.22c}$$

Then the same argument as in Lemma 4.4 from (4.31)-(4.34a, b) applies, and we have

$$\begin{aligned}&\phi_n(t)\to\tilde\phi(t) \quad \text{in } L^\infty(0,T;\mathscr{D}(A^{3/4})) \quad \text{weak star},\\ &\phi_n'(t)\to\tilde\phi'(t) \quad \text{in } L^\infty(0,T;\mathscr{D}(A^{1/4})) \quad \text{weak star},\end{aligned} \tag{6.23}$$

and thus $\phi_n(t)$ and $\phi_n'(t)$ are uniformly bounded in $L^\infty(0, T; \mathscr{D}(A^{3/4}))$ and $L^\infty(0, T; \mathscr{D}(A^{1/4}))$. We now use the compactness of the injections $\mathscr{D}(A^{3/4})\to\mathscr{D}(A^{1/2})$ and $\mathscr{D}(A^{1/4})\to L^2(\Omega)$ and recall (6.16) (while in the proof of Lemma 4.4 we used the compactness of the injection $\mathscr{D}(A^{3/4})\to\mathscr{D}(A^{1/4})=H_0^1(\Omega)$). As a consequence there is a subsequence, still subindexed by n, such that

$$\begin{aligned}&\phi_n(t)\to\tilde\phi(t) \quad \text{strongly in } L^\infty(0,T;\mathscr{D}(A^{1/2})),\\ &\phi_n'(t)\to\tilde\phi'(t) \quad \text{strongly in } L^\infty(0,T;L^2(\Omega))\end{aligned} \tag{6.24}$$

(counterpart of (4.35)). A fortiori, from (6.22), (6.24), and (6.16), we obtain for the sequence $\{\phi_n\}$ of solutions of (2.20a-d):

$$1\equiv\|\Delta\phi_n\|_{C([0,T];L^2(\Omega))}\to\|\Delta\tilde\phi\|_{C([0,T];L^2(\Omega))}=1, \tag{6.25}$$

$$1\equiv\|\phi_n'\|_{C([0,T];L^2(\Omega))}\to\|\tilde\phi'\|_{C([0,T];L^2(\Omega))}=1, \tag{6.26}$$

as well as (from (6.22c))

$$\left.\frac{\partial(\Delta\tilde\phi)}{\partial\nu}\right|_\Sigma\equiv 0 \quad \text{and} \quad \Delta\tilde\phi|_\Sigma\equiv 0. \tag{6.27}$$

Thus, $\tilde\phi(t)$ satisfies

$$\begin{aligned}&\tilde\phi_{tt}+\Delta^2\tilde\phi=0\\ &\tilde\phi|_\Sigma\equiv 0, \quad \left.\frac{\partial\tilde\phi}{\partial\nu}\right|_\Sigma\equiv 0 \qquad \text{from (4.33a) written for } \phi_n,\\ &\left.\frac{\partial(\Delta\tilde\phi)}{\partial\nu}\right|_\Sigma\equiv 0, \quad \Delta\tilde\phi|_\Sigma\equiv 0 \quad \text{from (6.27)}\end{aligned} \tag{6.28}$$

on $[0, T]$. Then, for any $T>0$ arbitrarily small, Holmgren's classical uniqueness theorem of Remark 1.4 implies

$$\tilde\phi\equiv 0 \quad \text{in } Q \tag{6.29}$$

and this contradicts (6.25)-(6.26). The proof of Lemma 6.2 is complete. □

As an immediate corollary of Lemma 6.2 and (6.20) we obtain Corollary 6.3.

COROLLARY 6.3. *For any $T>0$, we have*

$$C_{h,T,\varepsilon,\varepsilon'}\left\{\int_\Sigma \left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2+|\Delta\phi|^2\,d\Sigma\right\}\geqq[(\rho C_1-\varepsilon G_h C_1)T-\varepsilon' C_{2h}]\|\{\phi^0,\phi^1\}\|_Z^2. \tag{6.30}$$

Step (vii). We next "absorb" $\int_\Sigma(\Delta\phi)^2\,d\Sigma$ by $\int_\Sigma(\partial(\Delta\phi)/\partial\nu)^2\,d\Sigma$ by using assumption (1.24).

LEMMA 6.4. *Under the uniqueness property* (1.24), *inequality* (6.30) *implies that for any $T>T_u$, T_u as in* (1.24), *there is a constant $C_T>0$ such that*

$$\int_\Sigma|\Delta\phi|^2\,d\Sigma\leqq C_T\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma. \tag{6.31}$$

Proof. Suppose by contradiction that there is a sequence $\{\phi_n(t)\}$ of solutions to problem (2.20a–d), as in (4.28a–d), (4.29a, b), such that

$$\int_\Sigma|\Delta\phi_n|^2\,d\Sigma\equiv 1, \tag{6.32a}$$

$$\int_\Sigma\left(\frac{\partial(\Delta\phi_n)}{\partial\nu}\right)^2 d\Sigma\to 0. \tag{6.32b}$$

Then, by the preceding analysis, each solution $\phi_n(t)$ satisfies (6.30) and thus the uniform bound as in (4.31) holds true. The same argument as in (4.31)–(4.34) applies, so that from (4.34a) coupled with $\mathcal{D}(A^{3/4})=V$, V as in (1.4), we deduce that

$$\{A^{1/2}\phi_n\}\text{ is uniformly bounded in }L^\infty(0,T;\mathcal{D}(A^{1/4})). \tag{6.33}$$

Hence, by (2.32) and (2.33), we also have that

$$\{\Delta\phi_n\}\text{ is uniformly bounded in }L^\infty(0,T;\mathcal{D}(A^{1/4})). \tag{6.34}$$

By (2.32) and (4.29) we actually have $|\nabla(\Delta\phi_n)|\in C([0,T];L^2(\Omega))$ and $\Delta\phi_n\in C([0,T];L^2(\Omega))$. Thus, $\Delta\phi_n\in C([0,T];H^1(\Omega))$. By standard trace theory we deduce

$$\{\Delta\phi_n|_\Gamma\}\text{ uniformly bounded in }C([0,T];H^{1/2}(\Gamma)), \tag{6.35}$$

and thus

$$\{\Delta\phi_n|_\Gamma\}\text{ lies in a compact set of }C([0,T];L^2(\Gamma)). \tag{6.36}$$

Hence, recalling (4.34a, b) we obtain

$$\Delta\phi_n|_\Sigma\to\Delta\tilde{\phi}|_\Sigma\quad\text{in }L^2(0,T;L^2(\Gamma))\quad\text{strongly}. \tag{6.37}$$

Invoking (6.32a), we then conclude that

$$\int_\Sigma|\Delta\tilde{\phi}|^2\,d\Sigma=1. \tag{6.38}$$

On the other hand, (6.32b) implies

$$\left.\frac{\partial(\Delta\tilde{\phi})}{\partial\nu}\right|_{\Sigma} \equiv 0. \tag{6.39}$$

Thus $\tilde{\phi}$ satisfies

$$\begin{aligned} &\tilde{\phi}_{tt} + \Delta^2\tilde{\phi} \equiv 0 \\ &\tilde{\phi}|_{\Sigma} \equiv 0 \qquad \text{from (4.33)}, \\ &\left.\frac{\partial\tilde{\phi}}{\partial\nu}\right|_{\Sigma} \equiv 0, \\ &\left.\frac{\partial(\Delta\tilde{\phi})}{\partial\nu}\right|_{\Sigma} \equiv 0 \qquad \text{from (6.33)} \end{aligned} \tag{6.40}$$

for $0 \leqq t \leqq T_u < T$. The uniqueness property (1.24), which we have assumed, then implies that $\tilde{\phi} \equiv 0$ in Q, a contradiction to (6.38). The proof of Proposition 6.1 is complete. □

Remark 6.1. A (much simpler) modification of the arguments above leads to the sought-after exact controllability on X of problem (1.1a-d), (1.2) under different assumptions, whereby the undesirable assumption (1.24) on uniqueness is eliminated and replaced with another assumption on the vector field h, stating qualitatively that the constant G_h in (6.4) is small with respect to the constant ρ in (1.23).

To state this new result, we first introduce some constants:

$$\begin{aligned} &4G_h \equiv \max_{\bar{\Omega}} |\nabla(\operatorname{div} h)| \quad \text{from (6.4)} \\ &\cdot \int_{\Omega} \psi^2\, d\psi \leqq C_p \int_{\Omega} |\nabla\psi|^2\, d\Omega, \quad \psi \in H_0^1(\Omega), \quad C_p = \text{Poincaré constant} \end{aligned} \tag{6.41}$$

and according to the norm-equivalences (2.32), (2.33),

$$c\|A^{3/4}f\|_{\Omega}^2 \leqq \int_{\Omega} |\nabla(\Delta f)|^2\, d\Omega \leqq C\|A^{3/4}f\|_{\Omega}^2, \tag{6.42}$$

$$k\|A^{1/4}f\|_{\Omega}^2 \leqq \int_{\Omega} |\nabla f|^2\, d\Omega \leqq K\|A^{1/4}f\|_{\Omega}^2; \tag{6.43}$$

finally

$$C_m = \max\{\|A^{-1/4}\|, KC_p\}. \tag{6.44}$$

We can now state the following variation of Theorem 1.5, which extends Theorem 1.1 to the case where $G_h \neq 0$ (see [T1] for a similar result for the wave equation with Dirichlet boundary control).

THEOREM 6.2. *The conclusion of Theorem* 1.1 *holds true for problem* (1.1a-d)-(1.2), *with T sufficiently large as in* (6.51) *below, provided* Ω *satisfies the following geometrical conditions. There exists a vector field* $h(x) = [h_1(x), \cdots, h_n(x)] \in C^2(\bar{\Omega})$ *such that*

(i) $h \cdot \nu \geqq$ *constant* $\gamma > 0$ *on* Γ (*assumption* (1.21));

(ii) $\int_{\Omega} H(x)v(x) \cdot v(x)\, d\Omega \geqq \rho \int_{\Omega} |v(x)|_{R^n}^2\, d\Omega$, *for some* $\rho > 0$ *and all* $v \in [L^2(\Omega)]^n$ (*assumption* (1.23));

(iii) *With reference* (6.41)-(6.44),

$$\rho(c+k) - 2G_hC_m > 0. \tag{6.45}$$

Proof. Only the modifications on the proof of Theorem 1.5 will be indicated. We return to identity (6.2) and estimate as follows:

$$\begin{aligned}\frac{1}{2}\int_Q \Delta\phi\nabla(\operatorname{div} h)\cdot\nabla(\Delta\phi)\,dQ &\geqq -2G_h\int_Q |\Delta\phi||\nabla(\Delta\phi)|\,dQ\\ &\geqq -2G_h\|\Delta\phi\|_Q\|\|\nabla(\Delta\phi)\|\|_Q \quad \text{(by Schwarz's inequality)}\\ &\geqq -2G_hC\|A^{1/2}\phi\|_Q\|A^{3/4}\phi\|_Q \quad \text{(by (6.16) and (6.42))}\\ &\geqq -2G_hC\|A^{-1/4}\|\,\|A^{3/4}\phi\|_Q^2\end{aligned} \tag{6.46}$$

in place of (6.5). Similarly,

$$\begin{aligned}\frac{1}{2}\int_Q \phi_t\nabla(\operatorname{div} h)\cdot\nabla\phi_t\,dQ &\geqq -2G_h\int_Q |\phi_t||\nabla\phi_t|\,dQ \geqq -2G_h\|\phi_t\|_Q\|\|\nabla\phi_t\|\|_Q\\ &\geqq -2G_hC_p\|\|\nabla\phi_t\|\|_Q^2\\ &\geqq -2G_hC_pK\|A^{1/4}\phi_t\|_Q^2 \quad \text{(by (6.41) since } \phi_t\equiv \text{on } \Sigma)\end{aligned} \tag{6.47}$$

in place of (6.6). Thus, using (6.46), (6.47) in the RHS of identity (6.2) and recalling assumption (ii), we obtain

$$\begin{aligned}\text{RHS of (6.2)} &\geqq \rho\int_Q |\nabla(\Delta\phi)|^2+|\nabla\phi_t|^2\,dQ - 2G_hC\|A^{-1/4}\|\,\|A^{3/4}\phi\|_Q^2\\ &\quad -2G_hC_pK\|A^{1/4}\phi_t\|_Q^2+\beta_{0,T,h}\\ &\geqq [\rho(c+k)-2G_hC_m]\{\|A^{3/4}\phi\|_Q^2\\ &\quad +\|A^{1/4}\phi_t\|_Q^2\}+\beta_{0,T,h} \quad \text{(by (6.44))}.\end{aligned} \tag{6.48}$$

Thus, invoking the time invariance identity (2.29) and (2.39) as well, we obtain

$$\begin{aligned}\text{RHS of (6.2)} &\geqq [\rho(c+k)-2G_hC_m]T\{\|A^{3/4}\phi^0\|_\Omega^2+\|A^{1/4}\phi^1\|_\Omega^2\}\\ &\quad -\text{const}_{h,n}\{\|A^{3/4}\phi^0\|_\Omega^2+\|A^{1/4}\phi^1\|_\Omega^2\}.\end{aligned} \tag{6.49}$$

Moreover, for the LHS of (6.2) we have, as before,

$$C_{1nh\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma+\varepsilon c_h\int_0^T\|A^{3/4}\phi\|_\Omega^2\,dt \geqq \text{LHS of (6.2)} \tag{6.50}$$

(see (2.45), or else (6.13) combined with (2.42)).

Thus (6.50) and (6.49) combined yield (by proceeding as in obtaining (2.47) in particular, by using the time invariance (2.29))

$$C_{1nh\varepsilon}\int_\Sigma\left(\frac{\partial(\Delta\phi)}{\partial\nu}\right)^2 d\Sigma\geqq\{[\rho(c+k)-2G_hC_m-\varepsilon c_h]T-\text{const}_{h,n}\}\|\{\phi^0,\phi^1\}\|_Z^2$$

and the conclusion follows via Lemma 2.1(ii), (2.11) with

$$T>\frac{\text{const}_{h,n}}{\rho(c+k)-2G_hC_m}. \tag{6.51}$$

Appendix A. Proof of a general identity and of (2.22). For future reference to exact controllability problems for (1.1a) with boundary conditions of a type possibly different from (1.1c, d), we shall first derive a general identity for ϕ only solution of (2.20a) with no use of boundary conditions (2.20c, d), in terms of an arbitrary smooth vector field $h(x)\in C^2(\bar{\Omega})$ (see (A8) below). Only subsequently, we shall specialize such an

identity (A8) to ϕ that satisfies also the boundary conditions (2.20c-d), and to $h(x)$ that is a radial vector field $h(x) = x - x_0$, thus obtaining (2.22).

Identity for ϕ that satisfies (2.20a) for general vector field $h(x)$. Let $h(x) \in C^2(\bar{\Omega})$. With reference to Remark 1.4, we multiply (2.20a) by $h \cdot \nabla(\Delta\phi)$ and integrate over Q. We shall use the identity

$$\int_\Omega h \cdot \nabla f \, d\Omega = \int_\Gamma f h \cdot \nu \, d\Gamma - \int_\Omega f \operatorname{div} h \, d\Omega \tag{A1}$$

obtained from $\operatorname{div}(fh) = h \cdot \nabla f + f \operatorname{div} h$, f scalar function, and the divergence theorem. In addition, we shall use the identity

$$\begin{aligned} \int_Q \Delta\psi (h \cdot \nabla\psi) \, dQ = \int_\Sigma \frac{\partial\psi}{\partial\nu} (h \cdot \nabla\psi) \, d\Sigma - \frac{1}{2} \int_\Sigma |\nabla\psi|^2 h \cdot \nu \, d\Sigma \\ - \int_Q H\nabla\psi \cdot \nabla\psi \, dQ + \frac{1}{2} \int_Q |\nabla\psi|^2 \operatorname{div} h \, dQ, \end{aligned} \tag{A2}$$

already proved in, say, [T2, (A3) of Appendix A] (with similar multiplier $h \cdot \nabla\psi$), where $H = H(x)$ is the transpose of the Jacobian matrix of $h(x)$:

$$H(x) = \begin{vmatrix} \dfrac{\partial h_1}{\partial x_1} & \cdots & \dfrac{\partial h_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \dfrac{\partial h_n}{\partial x_1} & \cdots & \dfrac{\partial h_n}{\partial x_n} \end{vmatrix}. \tag{A3}$$

Term $\phi_{tt} h \cdot \nabla(\Delta\phi)$. Integrating at first by parts in t:

$$\int_\Omega \int_0^T \phi_{tt} h \cdot \nabla(\Delta\phi) \, dt \, d\Omega = \left[\int_\Omega \phi_t h \cdot \nabla(\Delta\phi) \, d\Omega \right]_0^T - \int_Q \phi_t h \cdot \nabla(\Delta\phi_t) \, dQ$$

(using (A1) with h replaced by $\phi_t h$, with $f = \Delta\phi_t$, and with $\operatorname{div}(\phi_t h) = \nabla\phi_t \cdot h + \phi_t \operatorname{div} h$)

$$\begin{aligned} = \left[\int_\Omega \phi_t h \cdot \nabla(\Delta\phi) \, d\Omega \right]_0^T - \int_\Sigma \phi_t \Delta\phi_t h \cdot \nu \, d\Sigma \\ + \int_Q \Delta\phi_t h \cdot \nabla\phi_t \, dQ + \int_Q \Delta\phi_t \phi_t \operatorname{div} h \, dQ. \end{aligned} \tag{A4}$$

If we use identity (A2) with $\psi = \phi_t$ for the third integral on the right of (A4),

$$\begin{aligned} \int_\Omega \int_0^T \phi_{tt} h \cdot \nabla(\Delta\phi) \, dt \, d\Omega = \left[\int_\Omega \phi_t h \cdot \nabla(\Delta\phi) \, d\Omega \right]_0^T - \int_\Sigma \phi_t \Delta\phi_t h \cdot \nu \, d\Sigma \\ - \frac{1}{2} \int_\Sigma |\nabla\phi_t|^2 h \cdot \nu \, d\Sigma + \int_\Sigma \frac{\partial\phi_t}{\partial\nu} h \cdot \nabla\phi_t \, d\Sigma \\ - \int_Q H\nabla\phi_t \cdot \nabla\phi_t \, dQ + \frac{1}{2} \int_Q |\nabla\phi_t|^2 \operatorname{div} h \, dQ \\ + \int_Q \Delta\phi_t \phi_t \operatorname{div} h \, dQ. \end{aligned} \tag{A5}$$

Using Green's first theorem on the last integral at the right of (A5) along with the identity

$$\nabla\phi_t \cdot \nabla(\phi_t \operatorname{div} h) = \phi_t \nabla(\operatorname{div} h) \cdot \nabla\phi_t + |\nabla\phi_t|^2 \operatorname{div} h,$$

we finally obtain from (A5)

$$\begin{aligned}\int_Q \phi_{tt} h\cdot\nabla(\Delta\phi)\,dQ &= [(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T - \int_\Sigma \phi_t\Delta\phi_t h\cdot\nu\,d\Sigma\\ &\quad -\frac{1}{2}\int_\Sigma |\nabla\phi_t|^2 h\cdot\nu\,d\Sigma + \int_\Sigma \frac{\partial\phi_t}{\partial\nu} h\cdot\nabla\phi_t\,d\Sigma\\ &\quad +\int_\Sigma \frac{\partial\phi_t}{\partial\nu}\phi_t \operatorname{div} h\,d\Sigma - \int_Q H\nabla\phi_t\cdot\nabla\phi_t\,dQ\\ &\quad -\frac{1}{2}\int_Q |\nabla\phi_t|^2 \operatorname{div} h\,dQ - \int_Q \phi_t\nabla(\operatorname{div} h)\cdot\nabla\phi_t\,dQ.\end{aligned} \tag{A6}$$

Term $\Delta^2\phi h\cdot\nabla(\Delta\phi)$. Using identity (A2), this time with $\psi=\Delta\phi$, we obtain

$$\begin{aligned}\int_Q \Delta(\Delta\phi)h\cdot\nabla(\Delta\phi)\,dQ &= \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla(\Delta\phi)\,d\Sigma - \frac{1}{2}\int_\Sigma |\nabla(\Delta\phi)|^2 h\cdot\nu\,d\Sigma\\ &\quad -\int_Q H\nabla(\Delta\phi)\cdot\nabla(\Delta\phi)\,dQ + \frac{1}{2}\int_Q |\nabla(\Delta\phi)|^2 \operatorname{div} h\,dQ.\end{aligned} \tag{A7}$$

Summing up (A6) and (A7), we finally obtain

$$\begin{aligned}&\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla(\Delta\phi)\,d\Sigma + \int_\Sigma \frac{\partial\phi_t}{\partial\nu} h\cdot\nabla\phi_t\,d\Sigma + \int_\Sigma \frac{\partial\phi_t}{\partial\nu}\phi_t \operatorname{div} h\,d\Sigma\\ &\quad -\frac{1}{2}\int_\Sigma |\nabla\phi_t|^2 h\cdot\nu\,d\Sigma - \frac{1}{2}\int_\Sigma |\nabla(\Delta\phi)|^2 h\cdot\nu\,d\Sigma - \int_\Sigma \phi_t\Delta\phi_t h\cdot\nu\,d\Sigma\\ &= \int_Q H\nabla(\Delta\phi)\cdot\nabla(\Delta\phi)\,dQ + \int_Q H\nabla\phi_t\cdot\nabla\phi_t\,dQ + \frac{1}{2}\int_Q \{|\nabla\phi_t|^2 - |\nabla(\Delta\phi)|^2\}\operatorname{div} h\,dQ\\ &\quad + \int_Q \phi_t\nabla(\operatorname{div} h)\cdot\nabla\phi_t\,dQ - [(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T,\end{aligned} \tag{A8}$$

which is the sought-after identity for ϕ satisfying (2.20a).

Specialization of left-hand side of (A8) to ϕ satisfying also the boundary conditions (2.20c, d). Recalling (2.20c, d) we have:

(A9a) $\phi_t|_\Sigma \equiv 0$, $\nabla\phi \perp \Gamma$ and $|\nabla\phi| = |\partial\phi/\partial\nu| \equiv 0$ on Σ by (2.20d),

(A9b) $\partial\phi_t/\partial\nu|_\Sigma \equiv 0$, $\nabla\phi_t \perp \Gamma$ and $|\nabla\phi_t| = |\partial\phi_t/\partial\nu| \equiv 0$ in Σ.

Thus, using (2.20c, d) and (A9a, b) in the LHS of (A8) we find that this simplifies to

$$\text{(A10)}\qquad \text{LHS of (A8)} = \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu} h\cdot\nabla(\Delta\phi)\,d\Sigma - \frac{1}{2}\int_\Sigma |\nabla(\Delta\phi)|^2 h\cdot\nu\,d\Sigma.$$

Specialization of the right-hand side of (A8) to radial vector fields $h(x)=x-x_0$. In this case, recalling (A3) we obtain

$$\text{(A11)}\qquad H(x)\equiv \text{identity matrix}, \qquad \operatorname{div} h \equiv n = \dim\Omega,$$

which used in the RHS of (A8) yield

$$\begin{aligned}\text{RHS of (A8)} &= \int_Q \{|\nabla\phi_t|^2 + |\nabla(\Delta\phi)|^2\}\,dQ\\ &\quad + \frac{n}{2}\int_Q \{|\nabla\phi_t|^2 - |\nabla(\Delta\phi)|^2\}\,dQ - [(\phi_t, h\cdot\nabla(\Delta\phi))_\Omega]_0^T.\end{aligned} \tag{A12}$$

Combining (A10) and (A12) proves (2.22), as desired.

Appendix B. Proof of identity (2.23). Again, we shall first obtain an identity, (B3) below, for ϕ that solves only (2.20a) and for an arbitrary smooth vector field $h \in C^2(\bar{\Omega})$. Next we shall specialize this identity (B3) to the case where ϕ satisfies, in addition, the boundary conditions (2.20c)-(2.20d) and, moreover, the vector field is radial.

We multiply (2.20a) by $\Delta\phi$ div h and integrate over Q by parts in t and by Green's first theorem:

(B1)

$$\begin{aligned}
&\int_\Omega \int_0^T \phi_{tt}\,\Delta\phi \operatorname{div} h\, dt\, d\Omega \\
&\quad = \left[\int_\Omega \Delta\phi\, \phi_t \operatorname{div} h\, d\Omega\right]_0^T - \int_0^T \int_\Omega \Delta\phi_t\, \phi_t \operatorname{div} h\, d\Omega\, dt \\
&\quad = \left[\int_\Gamma \frac{\partial\phi}{\partial\nu}\phi_t \operatorname{div} h\, d\Gamma - \int_\Omega \nabla\phi\cdot\nabla(\phi_t \operatorname{div} h)\, d\Omega\right]_0^T \\
&\qquad - \int_\Sigma \frac{\partial\phi_t}{\partial\nu}\phi_t \operatorname{div} h\, d\Sigma + \int_Q |\nabla\phi_t|^2 \operatorname{div} h\, dQ + \int_Q \phi_t \nabla(\operatorname{div} h)\cdot\nabla\phi_t\, dQ.
\end{aligned}$$

Also,

(B2)

$$\begin{aligned}
&\int_0^T \int_\Omega \Delta(\Delta\phi)\Delta\phi \operatorname{div} h\, d\Omega\, dt \\
&\quad = \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi \operatorname{div} h\, d\Sigma - \int_Q |\nabla(\Delta\phi)|^2 \operatorname{div} h\, dQ \\
&\qquad - \int_Q \Delta\phi \nabla(\operatorname{div} h)\cdot\nabla(\Delta\phi)\, dQ.
\end{aligned}$$

Summing up (B1) and (B2), we find the identity

(B3)

$$\begin{aligned}
&\int_Q \{|\nabla\phi_t|^2 - |\nabla(\Delta\phi)|^2\} \operatorname{div} h\, dQ \\
&\quad = \int_\Sigma \frac{\partial\phi_t}{\partial\nu}\phi_t \operatorname{div} h\, d\Sigma - \int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi \operatorname{div} h\, d\Sigma \\
&\qquad + \int_\Omega \Delta\phi \nabla(\operatorname{div} h)\cdot\nabla(\Delta\phi)\, dQ - \int_Q \phi_t \nabla(\operatorname{div} h)\cdot\nabla\phi_t\, dQ \\
&\qquad + \left[\int_\Omega \Delta\phi\cdot\nabla(\phi_t \operatorname{div} h)\, d\Omega - \int_\Gamma \frac{\partial\phi}{\partial\nu}\phi_t \operatorname{div} h\, d\Gamma\right]_0^T
\end{aligned}$$

for ϕ satisfying (2.20a).

Now if ϕ satisfies, in addition, the boundary conditions (2.20c, d), we have (for future use in § 5) the identity

(B4)

$$\begin{aligned}
&\int_Q \{|\nabla\phi_t|^2 - |\nabla(\Delta\phi)|^2\} \operatorname{div} h\, dQ \\
&\quad = -\int_\Sigma \frac{\partial(\Delta\phi)}{\partial\nu}\Delta\phi \operatorname{div} h\, d\Sigma + \int_Q \Delta\phi \nabla(\operatorname{div} h)\cdot\nabla(\Delta\phi)\, dQ \\
&\qquad - \int_Q \phi_t \nabla(\operatorname{div} h)\cdot\nabla\phi_t\, dQ + \left[\int_\Omega \nabla\phi\cdot\nabla(\phi_t \operatorname{div} h)\, d\Omega\right]_0^T.
\end{aligned}$$

Finally, if $h(x)$ is a radial vector field, then (B4) specializes to (2.23).

Appendix C. Proof of Lemma 2.3. We begin by recalling definition (2.2), the definition of the positive self-adjoint operator A. It suffices to show parts (i) and (ii), since part (iii) is a specialization of (ii) for ϕ solution of problem (2.20a-d).

Part (i). This follows from known interpolation results of Grisvard [G1] (see also Lions and Magenes [LM1]). If $0<\theta<1$, with $4\theta+\frac{1}{2}$ a nonpositive integer, then

$$\text{(C1)} \qquad \mathscr{D}(A^{\theta})=[\mathscr{D}(A), L^2(\Omega)]_{1-\theta}=\{f\in H^{4\theta}(\Omega): B_j f=0, \text{ if } m_j<4\theta-\tfrac{1}{2}\}$$

where the B_j are the boundary operators defining A and m_j their order. In our case we have $B_1=|_\Gamma$ and $B_2=\partial/\partial\nu|_\Gamma$ of orders zero and one, respectively. Then (C1) specializes at once to (2.30) and (2.31) for $\theta=\frac{3}{4}$ and $\theta=\frac{1}{4}$, respectively.

Part (ii). By (2.31), the $\mathscr{D}(A^{1/4})$-norm, defined by (1.6), is equivalent to the $H_0^1(\Omega)$-norm, which in turn is equivalent to the gradient norm by Poincaré inequality. This proves (2.33).

As for (2.32), we first note that $\int_\Omega |\nabla(\Delta f)|^2\, d\Omega=0$, i.e.,

$$\text{(C2)} \qquad \Delta f\equiv \text{const in } \Omega \quad \text{with } f|_\Gamma=\left.\frac{\partial f}{\partial \nu}\right|_\Gamma=0$$

readily implies $f\equiv 0$ in Ω. In fact, the boundary conditions give $|\nabla f|=|\partial f/\partial\nu|=0$ on Γ; hence $f=f_{x_j}\equiv 0$ on Γ. These, coupled with $\partial(\Delta f)/\partial x_j=\partial(\text{const})/\partial x_j$, i.e., $\Delta f_{x_j}=0$ in Ω, yield $f_{x_j}\equiv 0$ in Ω; hence $f=$ const in Ω and finally $f\equiv 0$ in Ω, as desired. Similarly, if $|\nabla(\Delta f)|\in L^2(\Omega)$, with $f=\partial f/\partial\nu=0$ on Γ, i.e., $f=f_{x_j}=0$ on Γ, then $f\in H^3(\Omega)$. In fact, $\partial(\Delta f)/\partial x_j=\Delta f_{x_j}\equiv v_j\in L^2(\Omega)$ and $f_{x_j}=0$ on Γ yield $f_{x_j}\in H^2(\Omega)$ by elliptic theory. This along with $f|_\Gamma=0$ yields $f\in L^2(\Omega)$ by Poincaré inequality; thus $f\in H^3(\Omega)$, as desired. □

Appendix D. The minimal norm steering control. Once exact controllability is established, the following elementary argument provides the minimal norm steering control u^0. We shall first carry out the reasoning for an abstract equation, and then specialize its conclusions as they apply to, say, problem (1.1a-d) with $\Gamma_0=\varnothing$, on the state space $H^1(\Omega)\times L^2(\Omega)$ with control space $L^2(\Sigma)$.

Abstract treatment. Consider the abstract equation

$$\text{(D1)} \qquad \dot{y}=\mathscr{A}y+\mathscr{B}u, \qquad y(0)=y_0,$$

$\mathscr{A}$ being the generator of an s.c. semigroup of operators on the Hilbert space Y and $\mathscr{B}: U\supset D(\mathscr{B})\to Y$ being a linear, generally unbounded operator from another Hilbert space U to Y, with $\mathscr{A}^{-1}\mathscr{B}$ continuous from U to Y (without loss of generality we may assume that $\mathscr{A}$ is boundedly invertible). The solution to (D1) with $y_0=0$ is

$$\text{(D2)} \qquad \mathscr{L}_T u=\mathscr{A}\int_0^T e^{\mathscr{A}(T-t)}\mathscr{A}^{-1}\mathscr{B}u(t)\, dt.$$

Let z be a target state in Y and consider the following minimization problem: Minimize

$$J(u)=\tfrac{1}{2}\|u\|^2_{L_2(0,T;U)}$$

over all $u\in L_2(0, T; U)$ such that $\mathscr{L}_T u=z$, under the (exact controllability) assumption that there exists at least one such u. If we indicate by $((\ ,\ ))$ the duality pairing between Y' and Y, the Lagrangian can be written as

$$L(u,p)=\tfrac{1}{2}(u,u)_{L_2(0,T;U)}-((p,\mathscr{L}_T u-z)), \qquad p\in Y'.$$

Taking $L_u=0$ yields

$$\text{(D3)} \qquad u^0=\mathscr{L}_T^{\#}p^0; \quad \text{thus } z=\mathscr{L}_T u^0=\mathscr{L}_T\mathscr{L}_T^{\#}p^0,$$

where $\mathscr{L}_T^\#$ is the conjugate operator from Y' to $L^2(0, T; U)$ defined by $((v, \mathscr{L}_T u)) = (\mathscr{L}_T^\# v, u)_{L^2(0,T;U)}$, $v \in Y'$. From (D2) it readily follows that

$$\mathscr{L}_T^\# p^0 = \mathscr{B}^* e^{\mathscr{A}^*(T-t)} J^{-1} p^0, \tag{D4}$$

where J is the norm-preserving isomorphism Y onto Y' given by Riesz's theorem. Note that

$$\mathscr{L}_T^* = \mathscr{L}_T^\# J \tag{D5}$$

is the Hilbert space adjoint $Y \to L^2(0, T; U)$ of $\mathscr{L}_T$ which we have used in the paper. Moreover, from (D3), (D5) we obtain

$$\begin{aligned} ((p^0, z)) &= ((p^0, \mathscr{L}_T \mathscr{L}_T^\# p^0)) = \|\mathscr{L}_T^\# p^0\|^2_{L^2(0,T;U)} = \|\mathscr{L}_T^* J^{-1} p^0\|^2_{L^2(0,T;U)} \\ &\geqq C_T \|J^{-1} p^0\|^2_Y = C_T \|p^0\|^2_{Y'}, \end{aligned} \tag{D6}$$

where we have used the lower-bound inequality for $\mathscr{L}_T^*$ which states that $\mathscr{L}_T$ is $L^2(0, T; U)$ onto Y, i.e., the exact controllability assumption on (D1). Thus, the operator $\mathscr{L}_T \mathscr{L}_T^\#$ defines an isomorphism Y' onto Y by the Lax-Milgram Theorem applied to (D6), and from (D3) we have

$$p^0 = [p_1^0, p_2^0] = (\mathscr{L}_T \mathscr{L}_T^\#)^{-1} z \in Y'. \tag{D7}$$

Hence, by (D3), (D4), and (D7) we find that the optimal minimal norm steering control is given by

$$u^0 = \mathscr{L}_T^\# (\mathscr{L}_T \mathscr{L}_T^\#)^{-1} z = \mathscr{B}^* e^{\mathscr{A}^*(T-t)} J^{-1} (\mathscr{L}_T \mathscr{L}_T^\#)^{-1} z \tag{D8}$$

in terms of the target z and the dynamics $\mathscr{A}$, $\mathscr{B}$, $\mathscr{L}_T$.

Specialization to problem (1.1a–d) with $U = U_1 \times U_2$, $U_1 = L^2(\Gamma)$, $U_2 = \{0\}$, $Y = [\mathscr{D}(A^{1/4})]' \times [\mathscr{D}(A^{3/4})]'$ so that $Y' = \mathscr{D}(A^{1/4}) \times \mathscr{D}(A^{3/4})$. Here the norm preserving isomorphism J from Y onto Y' is defined by $J = A^{-1/2} \times A^{-3/2}$ so that if $p^0 = [p_1^0, p_2^0] \in Y'$ then $J^{-1} p^0 = [A^{1/2} p_1^0, A^{3/2} p_2^0] \in Y$. Using the operator model for problem (1.1a–d) (see e.g., [FLT1, App. C, case 2]) we obtain that

$$u^0 = \mathscr{L}_T^\# p^0 = \mathscr{B}^* e^{\mathscr{A}^*(T-t)} J^{-1} p^0 = \frac{\partial \Delta \phi(t)}{\partial \nu}\bigg|_\Gamma \tag{D9}$$

where ϕ solves problem (2.9a–d) with

$$\begin{aligned} \phi^0 &= A^{-3/2} [J^{-1} p^0]_2 = p_2^0 \in \mathscr{D}(A^{3/4}), \\ \phi^1 &= -A^{-1/2} [J^{-1} p^0]_1 = -p_1^0 \in \mathscr{D}(A^{1/4}). \end{aligned} \tag{D10}$$

The case $U_1 = \{0\}$, $U_2 = L^2(\Gamma)$ with $Y = L^2(\Omega) \times H^{-2}(\Omega)$ can also be treated (see [FLT1, App. C, Case 1]) and leads to

$$u^0 = \mathscr{B}^* e^{\mathscr{A}^*(T-t)} J^{-1} p^0 = \Delta \phi(t)|_\Gamma \tag{D11]}$$

where ϕ solves problem (2.9a–d) with $\phi^0 = -p_2^0 \in H_0^2(\Omega)$ and $\phi^0 = p_1^0 \in L^2(\Omega)$. The minimal norm steering control (D11) is the one used in [L3] (through a different approach) in the case $g_1 = 0$, $g_2 \in L^2(\Sigma)$ for problem (1.1a–d), while the case $g_1 \in L^2(\Sigma)$, $g_2 = 0$ which was proposed for investigation in [L3] leads in fact to the minimal norm steering control (D9).

A more extended discussion of the conceptual content of this Appendix is given in [LT3, App. B].

REFERENCES

[BT1] J. BARTOLOMEO AND R. TRIGGIANI, *Uniform exponential energy decay of Euler-Bernoulli equations with feedback operators in the Dirichlet-Neumann boundary conditions*, 1988, to appear.

[BLR1] C. BARDOS, G. LEBEAU, AND R. RAUCH, *Contrôle et stabilisation de l'équation des ondes*, Publication de l'Ecole Polytechnique Palaiseau, Palaiseau, France, 1987.

[DR1] S. DOLECKI AND D. L. RUSSELL, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185-220.

[F1] A. FRIEDMAN, *Partial Differential Equations*, Holt, Reinhart, and Winston, New York, 1969.

[FLT1] F. FLANDOLI, I. LASIECKA, AND R. TRIGGIANI, *Algebraic Riccati equations with non-smoothing observation arising in hyperbolic and Euler-Bernoulli equations*, Ann. Mat. Pura Appl., to appear.

[G1] P. GRISVARD, *Caracterization de quelques espaces d'interpolation*, Arch. Rational Mech. Anal., 25 (1967), pp. 40-63.

[H1] L. F. HO, *Observabilité frontière de l'équation des ondes*, C.R. Acad. Sci. Paris, 302, 1986, pp. 443-446.

[H2] L. HORMANDER, *Linear Partial Differential Operators*, Springer-Verlag, Berlin, New York, 1969.

[K1] V. KOMORNIK, *Contrôlabilité exacte en un temps minimal*, C.R. Acad. Sci. Paris Sér. I Math., 304 (1987).

[LL1] J. LAGNESE AND J. L. LIONS, *Modeling, Analysis and Control of Thin Plates*, Masson, Paris, 1988.

[LLT1] 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.

[LT1] I. LASIECKA AND R. TRIGGIANI, *Uniform exponential energy decay of the wave equation in a bounded region with $L_2(0, \infty; L_2(\Gamma))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340-390.

[LT2] ———, *A cosine operator approach to modeling $L_2(0, T; L_2(\Gamma))$-boundary input hyperbolic equations*, Appl., Math. Optim., 7 (1981), pp. 35-83.

[LT3] ———, *Exact controllability for the wave equation with Neumann boundary control*, Appl. Math. Optim., to appear.

[LT4] ———, *Regularity of hyperbolic equations under $L_2(0, T; L_2(\Gamma))$-boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[LT5] ———, *Riccati equations for hyperbolic partial differential equations with $L_2(0, T; L_2(\Gamma))$-Dirichlet boundary terms*, SIAM J. Control Optim., 24 (1986), pp. 884-926.

[LT6] ———, *Exact controllability of the Euler-Bernoulli equation with $L^2(\Sigma)$-control only in the Dirichlet boundary conditions*, Atti Accad. Naz. Lincei; Rend. Cl. Sci. Fis. Rend. (13), LXXXII (1987).

[LT7] ———, *Regularity theory for a class of Euler-Bernoulli equations: a cosine operator approach*, Boll. Un. Mat. Ital., B(7) (1988).

[LT8] ———, *Exact controllability of the Euler-Bernoulli equation with boundary controls for displacement and moments*, J. Math. Anal. Appl., to appear.

[LT9] ———, *A direct approach to exact controllability for the wave equation with Neumann boundary control and to an Euler-Bernoulli equation*, in Proc. 26th Conference on Decision and Control, Los Angeles, 1987, pp. 52-534.

[L1] J. L. LIONS, *Contrôle des systèmes distribués singuliers*, Gauthier-Villars, Paris, 1983.

[L2] ———, *Un résultat de régularite pour l'opérateur $(\partial^2/\partial t) + \Delta^2$*, in Current Topics in Partial Differential Equations, Y. Ohya et al., eds., Kinokuniya Company, Tokyo, 1986.

[L3] ———, *Exact controllability, stabilization and perturbations*, SIAM Rev., 30 (1988), pp. 1-68.

[L4] ———, *Contrôlabilité exacte de systèmes distribués*, C.R. Acad. Sci. Sér. I Math., 302 (1986), pp. 471-475.

[L5] ———, *Contrôlabilité exacte de systèmes distribués: remarques sur la théorie générale et ses applications*, in Proc. Seventh Internat. Conference on Analysis and Optimization of Systems, Antibes, France, June 1986, Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin, New York, 1986, pp. 1-13.

[L6] ———, *Exact controllability of distributed systems. An introduction*, in Proc. 25th Conference on Decision and Control, Athens, Greece, December 1986.

[L7] W. LITTMANN, private communication, 1986.

[L8] ———, *Boundary control theory for beams and plates*, in Proc. 25th Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 2007-2009.

[L9] W. LITTMANN, *Near optimal time boundary controllability for a class of hyperbolic equations*, in Proc. IFIPS Working Conf., Gainesville, FL, 1986, Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin, New York, 1987, pp. 307-312.

[L10] J. L. LIONS, *Cônlrolabilité exacte, perturbations et stabilization de systèmes distribués*, Vols. 1 and 2, Masson, Paris, 1988.

[LM1] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, Vols. I and II, Springer-Verlag, Berlin, New York, 1972.

[T1] R. TRIGGIANI, *A cosine operator approach to modeling $L_2(0, T; L_2(\Gamma))$-boundary input problems for hyperbolic systems*, Proc. Eighth IFIPS Conference, Wurzburg, Federal Republic of Germany, 1977, Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin, New York, 1978, pp. 380-390.

[T2] ———, *Exact boundary controllability on $L_2(\Omega) \times H^{-1}(\Omega)$ for the wave equation with Dirichlet control acting on a portion of the boundary, and related problems*, in Distributed Parameter Systems, Lecture Notes in Control and Information Sci. 102, Springer-Verlag, Berlin, New York, 1987, pp. 292-332; Appl. Math. Optim., 8 (1988), pp. 241-277.

[T3] ———, *Wave equation on a bounded domain with boundary dissipation: an operator approach*, in Operator Methods for Optimal Control Problems, Sung J. Lees, ed., Lecture Notes in Pure and Applied Mathematics 108, Marcel Dekker, New York, 1987, pp. 283-310; Recent Advances in Communication and Control Theory, R. E. Kalman and G. I. Marchuk, eds., Optimization Software, New York, 1987, pp. 262-286; J. Math. Anal. Appl., to appear.

[TL1] A. TAYLOR AND D. LAY, *Introduction to Functional Analysis*, Second edition, John Wiley, New York, 1978.

[Z1] E. ZUAZUA, *Contrôlabilité exacte d'un modèle de plaques vibrantes en un temps arbitrairement petit*, C.R. Acad. Sci. Paris Sér. I Math., 804 (1987), pp. 173-176.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 374-388, March 1989

007

EXACT BOUNDARY CONTROLLABILITY OF MAXWELL'S EQUATIONS IN A GENERAL REGION*

JOHN E. LAGNESE†

Abstract. By the Hilbert uniqueness method, it is proved that the evolution of solutions of Maxwell's equations in a general region can be exactly controlled by means of currents flowing tangentially in the boundary of the region.

Key words. Maxwell's equations, exact controllability, boundary controllability

AMS(MOS) subject classifications. 93C20, 35L50

1. Introduction and problem formulation. Let Ω be a bounded, open, connected region in $\mathbb{R}^3$ with a smooth boundary Γ. We suppose that Ω is occupied by an electromagnetic medium of constant electric permittivity ε and constant magnetic permeability μ. We further assume that the electrical charge density ρ and the current density J in Ω are zero. Let $E(x, t)$ and $H(x, t)$ denote the electric field and magnetic field, respectively, at a point $x \in \Omega$ at time $t \geqq 0$. These satisfy Maxwell's equations

$$\varepsilon \frac{\partial E}{\partial t} - \operatorname{curl} H = 0, \quad \mu \frac{\partial H}{\partial t} + \operatorname{curl} E = 0, \quad \operatorname{div} E = \operatorname{div} H = 0 \quad \text{in } \Omega, \quad t > 0. \tag{1.1}$$

It is assumed that the time evolution of the electric and magnetic fields is driven by an externally applied density of current flowing tangentially in Γ. We then have the boundary condition

$$\nu \times H = -J \quad \text{on } \Gamma, \qquad t > 0, \tag{1.2}$$

where ν is the unit normal vector to Γ pointing into the exterior of Ω. Let E^0, H^0 denote the distribution of E and H, respectively, at time $t = 0$:

$$E(0) = E^0, \qquad H(0) = H^0 \quad \text{in } \Omega. \tag{1.3}$$

In this paper we consider the following problem.

EXACT CONTROLLABILITY PROBLEM. Given the initial distribution $\{H^0, E^0\}$, a time $T > 0$, and a desired terminal state $\{H^T, E^T\}$ with $\{H^0, E^0\}, \{H^T, E^T\}$ in appropriate function spaces, find (if possible) a surface density of current J in a suitable function space such that the solution of (1.1)–(1.3) satisfies

$$E(T) = E^T, \qquad H(T) = H^T \quad \text{in } \Omega.$$

Remark 1.1. Because solutions of (1.1) propagate with finite velocity, the exact controllability problem can have a solution only if T is sufficiently large. The determination of T is part of the problem.

The exact controllability problem for Maxwell's equations with boundary control has been studied previously by Russell [6] for a right circular cylindrical region Ω, and by Kime [2] for a spherical region Ω. In the cylindrical region situation, it is assumed that the fields E and H (and control J) do not depend on the axial coordinate. The control problem can then be transformed into a problem of *simultaneous* exact controllability of a pair of ordinary wave equations in a circular region (i.e., the exact

* Received by the editors January 25, 1988; accepted for publication May 31, 1988. This research was supported by Air Force Office of Scientific Research grant AFOSR-86-0162.

† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

controllability of two wave equations by means of a single control). The latter control problem is then solved by the moment problem method. A moment problem approach is also used in [2] to find a solution to the exact controllability problem in a spherical region. In both [2] and [6], $\mathscr{L}^2$ boundary controls J are employed. It should be noted that the control functions J considered in [6] are further constrained by a *preassigned direction* in the tangent space to Γ, a very important requirement for applications. The control functions are not so constrained in [2] nor will they be in this paper. (In fact, for a general region, it is a challenge to even formulate such a constraint in a reasonable way.)

The moment problem approach employed in [2] and [6] is not feasible for dealing with the exact controllability problem in a general region. Rather, we shall attack the problem by means of the *Hilbert uniqueness method* (HUM) introduced by Lions [4], [5]. In this method we consider, in addition to the system (1.1), (1.2), the homogeneous adjoint system

$$\varepsilon\frac{\partial\varphi}{\partial t}-\operatorname{curl}\psi=0,\quad \mu\frac{\partial\psi}{\partial t}+\operatorname{curl}\varphi=0,\quad \operatorname{div}\varphi=\operatorname{div}\psi=0\quad\text{in }\Omega,\quad t>0, \tag{1.4}$$

$$\nu\times\varphi=0\quad\text{on }\Gamma,\qquad t>0. \tag{1.5}$$

Solvability of the exact controllability problem is shown to be equivalent to uniqueness of solutions of (1.4), (1.5) when certain additional *boundary data* are prescribed. We then construct Hilbert spaces associated with these uniqueness results (this can be done in infinitely many ways) and prove exact controllability in the duals of these spaces. In this way we obtain a variety of exact controllability results, depending on the particular Hilbert spaces constructed. In all cases, the method is *constructive*: the control J is defined in terms of the solution to (1.4), (1.5) with initial data

$$\varphi(0)=\varphi^0,\qquad \psi(0)=\psi^0\quad\text{in }\Omega, \tag{1.6}$$

where φ, ψ^0 are uniquely determined from the data of the original control problem.

The remainder of this paper is organized as follows. Section 2 defines the appropriate function spaces and discusses the well-posedness of problem (1.4)–(1.6). Energy estimates for solutions of (1.4), (1.5) are derived in § 3 (Lemmas 3.1–3.4); these are the basis for the application of HUM to our control problem. Three different exact controllability results are presented in Theorems 4.1–4.3 of § 4. They are distinguished by what we assume regarding regularity of the data—the geometry of Γ and the regularity that the control J possesses. In particular, it is proved that the system (1.1)–(1.3) is exactly controllable in the space $J(\Omega)\times\hat{J}(\Omega)$ (see (2.1), (2.2) below) using $\mathscr{L}^2(\Gamma\times(0,T))$ controls, provided Γ is star-shaped with respect to some point $x_0\in\Omega$ and T is $O(\sqrt{\varepsilon\mu})$ (this is made precise in (3.15) below).

2. Function spaces and well-posedness of (1.4)–(1.6). The spaces $L^2(\Omega)$, $L^2(\Gamma)$, $H^k(\Omega)$, $H^k(\Gamma)$ will denote the standard real L^2 and Sobolev spaces over Ω or Γ as notation implies. We shall use script notation to denote the corresponding spaces of $\mathbb{R}^3$-valued functions:

$$\mathscr{L}^2(\Omega)=(L^2(\Omega))^3,\qquad \mathscr{L}^2(\Gamma)=(L^2(\Gamma))^3,$$

$$\mathscr{H}^k(\Omega)=(H^k(\Omega))^3,\qquad \mathscr{H}^k(\Gamma)=(H^k(\Gamma))^3,$$

with the product topology in each case. The inner product and norm, respectively, in $\mathscr{L}^2(\Omega)$ are denoted by $(\cdot,\cdot)$ and $\|\cdot\|$.

We also introduce the following spaces (the notation is adopted from Ladyzhenskaya and Solonikov [3]):

$$J(\Omega) = \text{closure in } \mathscr{L}^2(\Omega) \text{ of } \{\chi | \chi \in C^\infty(\bar{\Omega}), \operatorname{div} \chi = 0\}, \tag{2.1}$$

$$\hat{J}(\Omega) = \text{closure in } \mathscr{L}^2(\Omega) \text{ of } \{\chi | \chi \in C_0^\infty(\Omega), \operatorname{div} \chi = 0\}, \tag{2.2}$$

where C^∞ denotes the class of infinitely differentiable $\mathbb{R}^3$-valued functions. For $k \geqq 1$, we set

$$J^k(\Omega) = J(\Omega) \cap \mathscr{H}^k(\Omega),$$

$$J_\nu^k(\Omega) = \{\chi | \chi \in J^k(\Omega), \nu \cdot \chi = 0 \text{ on } \Gamma\},$$

$$J_\tau^k(\Omega) = \{\chi | \chi \in J^k(\Omega), \nu \times \chi = 0 \text{ on } \Gamma\}$$

with the topology in each case that induced by $\mathscr{H}^k(\Omega)$. We further introduce

$$J_\nu^*(\Omega) = \{\chi | \chi \in J_\nu^2(\Omega), \nu \times \operatorname{curl} \chi = 0 \text{ on } \Gamma\},$$

$$J_\tau^*(\Omega) = \{\chi | \chi \in J_\tau^2(\Omega), \nu \cdot \operatorname{curl} \chi = 0 \text{ on } \Gamma\}$$

with topology in each space inherited from $\mathscr{H}^2(\Omega)$. The spaces above are known to have the following properties (see [3, § 7]):

$$J_\tau^*(\Omega) \subset J_\tau^1(\Omega) \subset J(\Omega), \qquad J_\nu^*(\Omega) \subset J_\nu^1(\Omega) \subset \hat{J}(\Omega) \tag{2.3}$$

with each space dense and continuously embedded in the one that follows it. In addition,

(2.4) If $k \geqq 1$, then the map $\varphi \to \operatorname{curl} \varphi$ is a continuous linear bijection of $J_\tau^k(\Omega)$ (respectively, $J_\nu^k(\Omega)$) onto $J_\nu^{k-1}(\Omega)$ (respectively, $J_\tau^{k-1}(\Omega)$), where $J_\tau^0(\Omega) \doteq J(\Omega)$ and $J_\nu^0(\Omega) \doteq \hat{J}(\Omega)$.

It follows from (2.3), (2.4) that $J_\tau^1(\Omega)$ and $J_\nu^1(\Omega)$ may be renormed using

$$\|\varphi\|_{J_\tau^1(\Omega)} = \|\operatorname{curl} \varphi\|, \qquad \|\psi\|_{J_\nu^1(\Omega)} = \|\operatorname{curl} \psi\|, \tag{2.5}$$

and that the norms (2.5) are *equivalent* to the $\mathscr{H}^1$ norms on these spaces.

Remark 2.1. If $\varphi \in J_\tau^1(\Omega)$ and $\psi \in J_\nu^1(\Omega)$, we have

$$(\operatorname{curl} \varphi, \psi) = (\varphi, \operatorname{curl} \psi);$$

hence

$$|(\operatorname{curl} \varphi, \psi)| \leqq \|\varphi\| \, \|\psi\|_{J_\nu^1(\Omega)}.$$

Therefore the mapping $\varphi \to \operatorname{curl} \varphi \colon J_\tau^1(\Omega) \to \hat{J}(\Omega)$ may be extended to a continuous linear mapping from $J(\Omega)$ into $(J_\nu^1(\Omega))'$, where $(J_\nu^1(\Omega))'$ denotes the dual of $J_\nu^1(\Omega)$ with respect to $\hat{J}(\Omega)$. Similarly, the mapping $\varphi \to \operatorname{curl} \varphi \colon J_\nu^1(\Omega) \to J(\Omega)$ has an extension to a continuous linear mapping from $\hat{J}(\Omega)$ into $(J_\tau^1(\Omega))'$, the dual of $J_\tau^1(\Omega)$ with respect to $J(\Omega)$. From (2.4) we may see that each of these extensions is a homeomorphism.

We now consider the well-posedness of the problem (1.4)–(1.6). Proceeding formally for the moment, let $\{\varphi, \psi\}$ be a solution. Since $\operatorname{div} \varphi = 0$, there is an $\mathbb{R}^3$-valued function W, determined up to a gradient ∇f, such that $\varepsilon\varphi = \operatorname{curl} W$. From the equation

$$0 = -\operatorname{curl} \psi + \varepsilon\varphi' = -\operatorname{curl} \psi + \operatorname{curl} W',$$

there is a real function g such that

$$\psi = W' + \nabla g.$$

It is classical that the function f may be chosen so that $g = 0$ and that, with this choice of f, W satisfies (see, e.g., Friedrichs [1])

$$\varepsilon\mu W'' + \operatorname{curl}(\operatorname{curl} W) = 0, \quad \operatorname{div} W = 0 \quad \text{in } \Omega, \quad t > 0, \tag{2.6}$$

$$\nu \times \operatorname{curl} W = 0 \quad \text{on } \Gamma, \qquad t > 0, \tag{2.7}$$

$$W(0) = W^0, \qquad W'(0) = \psi^0 \tag{2.8}$$

where $\varepsilon \operatorname{curl} \varphi^0 = W^0$.

Conversely, if W is a solution to (2.6)-(2.8), then setting $\varepsilon\varphi = \operatorname{curl} W$, $\psi = W'$, we see that $\{\varphi, \psi\}$ is a solution to (1.4), (1.5) with initial data $\varphi^0 = (1/\varepsilon) \operatorname{curl} W^0$, $\psi^0 = W^1$.

Now let us make the correspondence between (1.4)-(1.6) and (2.6)-(2.8) precise. First, we take the following variational problem as the *definition* of the problem (2.6)-(2.8):

$$\varepsilon\mu(W'', \hat{W}) + (\operatorname{curl} W, \operatorname{curl} \hat{W}) = 0 \quad \forall \hat{W} \in J^1_\nu(\Omega), \qquad t > 0, \tag{2.9}$$

$$W(0) = W^0 \in J^1_\nu(\Omega), \qquad W'(0) = W^1 \in \hat{J}(\Omega). \tag{2.10}$$

This as justified by the fact that (2.6)-(2.8) and (2.9), (2.10) are equivalent for smooth $J^1_\nu(\Omega)$-valued functions W. Next, we note that the form $a(W, \hat{W}) = (\operatorname{curl} W, \operatorname{curl} \hat{W})$ is strictly coercive on $J^1_\nu(\Omega)$. Since $J^1_\nu(\Omega)$ is dense in $\hat{J}(\Omega)$ with continuous injection, it follows from standard variational theory that there is one and only one function W satisfying

$$W \in C([0, \infty); J^1_\nu(\Omega)), \quad W' \in C([0, \infty); \hat{J}(\Omega)), \quad W'' \in C([0, \infty); (J^1_\nu(\Omega))')$$

and (2.9), (2.10). In (2.9), $(W'', \hat{W})$ is interpreted in the $(J^1_\nu(\Omega))' - J^1_\nu(\Omega)$ duality.

Remark 2.2. In view of Remark 2.1, (2.9) is equivalent to

$$\varepsilon\mu W'' + \operatorname{curl}(\operatorname{curl} W) = 0 \quad \text{in } (J^1_\nu(\Omega))'. \tag{2.11}$$

2.1. Weak solutions of (1.4)–(1.6). Suppose that $\{W^0, W^1\}$ satisfies (2.10) and let W be the solution to (2.9), (2.10). Define

$$\varepsilon\varphi = \operatorname{curl} W, \qquad \psi = W'. \tag{2.12}$$

Then

$$\begin{aligned} &\varphi \in C([0, \infty); J(\Omega)), \qquad \varphi' \in C([0, \infty); (J^1_\tau(\Omega))'), \\ &\psi \in C([0, \infty); \hat{J}(\Omega)), \qquad \psi' \in C([0, \infty); (J^1_\nu(\Omega))'), \\ &\varphi(0) = \varphi^0 \doteq (1/\varepsilon) \operatorname{curl} W^0 \in J(\Omega), \qquad \psi(0) = \psi^0 \doteq W^1 \in \hat{J}(\Omega), \end{aligned} \tag{2.13}$$

$$\mu\psi' + \operatorname{curl} \varphi = 0 \quad \text{in } (J^1_\nu(\Omega))'. \tag{2.14}$$

From (2.12) it is clear that

$$\varepsilon\varphi' - \operatorname{curl} \psi = 0 \quad \text{in } (J^1_\tau(\Omega))'. \tag{2.15}$$

Conversely, suppose that $\{\varphi, \psi\}$ satisfies (2.13)-(2.15) and

$$\varphi(0) = \varphi^0 \in J(\Omega), \qquad \psi(0) = \psi^0 \in \hat{J}(\Omega). \tag{2.16}$$

From (2.4) there is a unique function $W \in C([0, \infty); J^1_\nu(\Omega))$ such that $\operatorname{curl} W = \varepsilon\varphi$. Since, from (2.14), $\operatorname{curl} \psi = \varepsilon\varphi' = \operatorname{curl} W'$ in $(J^1_\tau(\Omega))'$, it follows from Remark 2.1 that $\psi = W'$. Therefore W satisfies (2.11), and

$$W(0) = W^0, \qquad W'(0) = \psi^0,$$

where $W^0 \in J^1_\nu(\Omega)$ is defined by $\operatorname{curl} W^0 = \varepsilon\varphi^0$.

We have therefore proved Theorem 2.1.

THEOREM 2.1. *Assume that* $\{\varphi^0, \psi^0\} \in J(\Omega) \times \hat{J}(\Omega)$. *There is one and only one pair* $\{\varphi, \psi\}$ *that satisfies* (2.13)-(2.16).

Remark 2.3 (*Conservation of energy for weak solutions*). From (2.9) it follows that

$$\varepsilon\mu \|W'(t)\|^2 + \|\text{curl } W(t)\|^2 = \varepsilon\mu \|W^1\|^2 + \|\text{curl } W^0\|^2. \tag{2.17}$$

In terms of φ, ψ, this conservation law is

$$\mu \|\psi(t)\|^2 + \varepsilon \|\varphi(t)\|^2 = \mu \|\psi^0\|^2 + \varepsilon \|\varphi^0\|^2. \tag{2.18}$$

2.2. Strong solutions of (1.4)–(1.6). Let us set $U = W'$ and formally differentiate (2.11) in t. Then U satisfies

$$\varepsilon\mu U'' + \text{curl}\,(\text{curl } U) = 0 \quad \text{in } (J^1_\nu(\Omega))', \tag{2.19}$$

$$U(0) = W^1, \qquad U'(0) = -(1/\varepsilon\mu)\,\text{curl}\,(\text{curl } W^0). \tag{2.20}$$

The system (2.19), (2.20) has a unique solution if

$$W^1 \in J^1_\nu(\Omega), \qquad \text{curl}\,(\text{curl } W^0) \in \hat{J}(\Omega),$$

that is, if curl $W^0 \in J^1_\tau(\Omega)$. It is standard theory that this solution is exactly W'; hence we have

$$W' \in C([0,\infty); J^1_\nu(\Omega)), \quad W'' \in C([0,\infty); \hat{J}(\Omega)), \quad \text{curl } W \in C([0,\infty); J^1_\tau(\Omega)). \tag{2.21}$$

It follows from (2.21) and the discussion leading to Theorem 2.1 that we have the following theorem.

THEOREM 2.2. *Assume that* $\{\varphi^0, \psi^0\} \in J^1_\tau(\Omega) \times J^1_\nu(\Omega)$. *There is one and only one pair* $\{\varphi, \psi\}$ *satisfying*

$$\{\varphi, \psi\} \in C([0,\infty); J^1_\tau(\Omega) \times J^1_\nu(\Omega)), \qquad \{\varphi', \psi'\} \in C([0,\infty); J(\Omega) \times \hat{J}(\Omega))$$

and (1.4)-(1.6).

Remark 2.4 (*Conservation of energy for strong solutions*). From (2.17), applied to $U = W'$, we have the conservation law

$$\varepsilon \|\text{curl } \varphi(t)\|^2 + \mu \|\text{curl } \psi(t)\|^2 = \varepsilon \|\text{curl } \varphi^0\|^2 + \mu \|\text{curl } \psi^0\|^2. \tag{2.22}$$

Let us take Theorem 2.2 a step further by differentiating (2.19) in t. Setting $V = U'$, we obtain

$$\varepsilon\mu V'' + \text{curl}\,(\text{curl } V) = 0 \quad \text{in } (J^1_\nu(\Omega))',$$

$$V(0) = -(1/\varepsilon\mu)\,\text{curl}\,(\text{curl } W^0), \qquad V'(0) = -(1/\varepsilon\mu)\,\text{curl}\,(\text{curl } W^1).$$

This problem is uniquely solvable, and the solution is exactly W'', provided

$$\text{curl}\,(\text{curl } W^0) \in J^1_\nu(\Omega), \qquad -(1/\varepsilon\mu)\,\text{curl}\,(\text{curl } W^1) \in \hat{J}(\Omega). \tag{2.23}$$

In terms of $\{\varphi^0, \psi^0\}$, (2.23) is curl $\varphi^0 \in J^1_\nu(\Omega)$, curl $\psi^0 \in J^1_\tau(\Omega)$, that is,

$$\varphi^0 \in J^*_\tau(\Omega), \qquad \psi^0 \in J^*_\nu(\Omega). \tag{2.24}$$

With (2.24), the solution of Theorem 2.2 satisfies

$$\varphi \in C([0,\infty); J^*_\tau(\Omega)), \quad \varphi' \in C([0,\infty); J^1_\tau(\Omega)), \quad \varphi'' \in C([0,\infty); J(\Omega)), \tag{2.25}$$

$$\psi \in C([0,\infty); J^*_\nu(\Omega)), \quad \psi' \in C([0,\infty); J^1_\nu(\Omega)), \quad \psi'' \in C([0,\infty); \hat{J}(\Omega)). \tag{2.26}$$

THEOREM 2.3. *Assume that* $\{\varphi^0, \psi^0\} \in J^*_\tau(\Omega) \times \hat{J}^*_\nu(\Omega)$. *Then the unique solution* $\{\varphi, \psi\}$ *of* (1.4)-(1.6) *satisfies* (2.25)-(2.26).

3. Energy estimates for solutions of the homogeneous problem. In this section we will derive energy estimates for solutions of (1.4), (1.5), that will be the basis for the application of HUM to the exact controllability problem.

We will assume that $\varepsilon = 1$ in (1.4), which amounts to the change $t \to t/\varepsilon$ in the time scale. (The reverse transformation will be done at the end.) Thus we consider the system

$$\frac{\partial \varphi}{\partial t} - \operatorname{curl} \psi = 0, \tag{3.1a}$$

$$\gamma \frac{\partial \psi}{\partial t} + \operatorname{curl} \varphi = 0, \qquad \operatorname{div} \varphi = \operatorname{div} \psi = 0 \quad \text{in } Q = \Omega \times (0, T), \tag{3.1b}$$

$$\nu \times \varphi = 0 \quad \text{on } \Sigma = \Gamma \times (0, T), \tag{3.2}$$

where $T > 0$ and $\gamma = \mu/\varepsilon$. The initial values are assumed to satisfy

$$\varphi(0) = \varphi^0 \in J_\tau^*(\Omega), \qquad \psi(0) = \psi^0 \in J_\nu^0(\Omega). \tag{3.3}$$

The solution to (3.1)–(3.3) then satisfies

$$\varphi \in C([0, T]; J_\tau^*(\Omega)), \quad \varphi' \in C([0, T]; J_\tau^1(\Omega)), \quad \varphi'' \in C([0, T]; J(\Omega)),$$
$$\psi \in C([0, T]; J_\nu^*(\Omega)), \quad \psi' \in C([0, T]; J_\nu^1(\Omega)), \quad \psi'' \in C([0, T]; \hat{J}(\Omega)).$$

We introduce the vector field m in $\mathbb{R}^3$ defined by

$$m(x; x_0) = x - x_0,$$

where x_0 is fixed. Let us form the inner product of (3.1a) and curl $(m \cdot \nabla)\psi$ and integrate the result over $\Omega \times (0, T)$. Thus,

$$\int_0^T (\varphi' - \operatorname{curl} \psi, \operatorname{curl} (m \cdot \nabla)\psi)\, dt = 0. \tag{3.4}$$

We have

$$\begin{aligned} \int_0^T (\varphi', \operatorname{curl} (m \cdot \nabla)\psi)\, dt &= \int_0^T (\operatorname{curl} \varphi', (m \cdot \nabla)\psi)\, dt + \int_\Sigma \nu \cdot (((m \cdot \nabla)\psi) \times \varphi')\, d\Gamma\, dt \\ &= -\gamma \int_0^T (\psi'', (m \cdot \nabla)\psi)\, dt \\ &= -\gamma (\psi', (m \cdot \nabla)\psi)\Big|_0^T + \gamma \int_0^T (\psi', (m \cdot \nabla)\psi')\, dt, \end{aligned} \tag{3.5}$$

since $\nu \times \varphi' = 0$ on Σ. We set

$$X_1 = -\gamma (\psi', (m \cdot \nabla)\psi)\Big|_0^T = (\operatorname{curl} \varphi, (m \cdot \nabla)\psi)\Big|_0^T.$$

Then (3.5) may be written:

$$\begin{aligned} \int_0^T (\varphi', \operatorname{curl} (m \cdot \nabla)\psi)\, dt &= X_1 + \frac{\gamma}{2} \int_Q \operatorname{div} (m|\psi'|^2)\, dx\, dt - \frac{3\gamma}{2} \int_Q |\psi'|^2\, dx\, dt \\ &= X_1 + \frac{\gamma}{2} \int_\Sigma m \cdot \nu |\psi'|^2\, d\Gamma\, dt - \frac{3\gamma}{2} \int_Q |\psi'|^2\, dx\, dt. \end{aligned} \tag{3.6}$$

Next, we note that

$$\operatorname{curl} (m \cdot \nabla)\psi = 2 \operatorname{curl} \psi + (m \cdot \nabla) \operatorname{curl} \psi - \operatorname{curl} ((\psi \cdot \nabla)m).$$

But $(\psi \cdot \nabla)m = \psi$, so that

$$\begin{aligned}(\operatorname{curl}\psi)\cdot(\operatorname{curl}(m\cdot\nabla)\psi) &= |\operatorname{curl}\psi|^2 + (\operatorname{curl}\psi)\cdot((m\cdot\nabla)\operatorname{curl}\psi)\\ &= |\operatorname{curl}\psi|^2 + \tfrac{1}{2}\operatorname{div}(m|\operatorname{curl}\psi|^2) - \tfrac{3}{2}|\operatorname{curl}\psi|^2\\ &= -\tfrac{1}{2}|\operatorname{curl}\psi|^2 + \tfrac{1}{2}\operatorname{div}(m|\operatorname{curl}\psi|^2).\end{aligned}$$

Therefore

(3.7)

$$\int_0^T (\operatorname{curl}\psi, \operatorname{curl}(m\cdot\nabla)\psi)\,dt = -\frac{1}{2}\int_Q |\operatorname{curl}\psi|^2\,dx\,dt + \frac{1}{2}\int_\Sigma m\cdot\nu|\operatorname{curl}\psi|^2\,d\Gamma\,dt.$$

Substituting (3.6), (3.7) into (3.4), we obtain

(3.8)

$$X_1 - \frac{3\gamma}{2}\int_Q |\psi'|^2\,dx\,dt + \frac{1}{2}\int_Q |\operatorname{curl}\psi|^2\,dx\,dt + \frac{1}{2}\int_\Sigma m\cdot\nu(\gamma|\psi'|^2 - |\operatorname{curl}\psi|^2)\,d\Gamma\,dt = 0.$$

Next, we form the inner product of (3.1b) with ψ' and integrate over $\Omega\times(0, T)$. We obtain

$$\begin{aligned}0 &= \gamma\int_Q |\psi'|^2\,dx\,dt + \int_0^T (\operatorname{curl}\varphi, \psi')\,dt\\ &= \gamma\int_Q |\psi'|^2\,dx\,dt + \int_0^T (\varphi, \operatorname{curl}\psi')\,dt + \int_\Sigma \nu\cdot(\varphi\times\psi')\,d\Gamma\,dt\\ &= \gamma\int_Q |\psi'|^2\,dx\,dt + \int_0^T (\varphi, \varphi'')\,dx\,dt + \int_\Sigma \psi'\cdot(\nu\times\varphi)\,d\Gamma\,dt\\ &= \gamma\int_Q |\psi'|^2\,dx\,dt + (\varphi, \varphi')|_0^T - \int_Q |\varphi'|^2\,dx\,dt.\end{aligned}$$

Therefore

$$\gamma\int_Q |\psi'|^2\,dx\,dt = \int_Q |\operatorname{curl}\psi|^2\,dx\,dt - X_2, \tag{3.9}$$

where

$$X_2 = (\varphi, \varphi')|_0^T = (\varphi, \operatorname{curl}\psi)|_0^T.$$

Substitution of (3.9) into (3.8) yields

$$\int_Q (\gamma|\psi'|^2 + |\operatorname{curl}\psi|^2)\,dx\,dt = 2(X_1 + X_2) + \int_\Sigma m\cdot\nu(\gamma|\psi'|^2 - |\operatorname{curl}\psi|^2)\,d\Gamma\,dt,$$

which may be written

$$\begin{aligned}&\int_Q (\varepsilon|\operatorname{curl}\varphi|^2 + \mu|\operatorname{curl}\psi|^2)\,dx\,dt\\ &\qquad = 2\mu(X_1 + X_2) + \mu\int_\Sigma m\cdot\nu(\gamma|\psi'|^2 - |\operatorname{curl}\psi|^2)\,d\Gamma\,dt.\end{aligned} \tag{3.10}$$

If we replace T by T/ε and rewrite (3.10) in the original time scale, we obtain

$$\int_Q (\varepsilon|\text{curl } \varphi|^2+\mu|\text{curl } \psi|^2)\, dx\, dt$$

$$(3.11)\qquad = 2\varepsilon\mu(X_1+X_2)+\mu\int_\Sigma m\cdot\nu(\mu\varepsilon|\psi'|^2-|\text{curl } \psi|^2)\, d\Gamma\, dt.$$

But since

$$\frac{d}{dt}\int_\Omega (\varepsilon|\text{curl } \varphi|^2+\mu|\text{curl } \psi|^2)\, dx = 0,$$

(3.11) becomes

$$T(\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2)$$

$$(3.12)\qquad = 2\varepsilon\mu(X_1+X_2)+\mu\int_\Sigma m\cdot\nu(\varepsilon\mu|\psi'|^2-|\text{curl } \psi|^2)\, d\Gamma\, dt.$$

3.1. A priori estimates. Let c_1, c_2 be the smallest constants such that

$$\|\chi\|_{\mathscr{H}^1(\Omega)} \leqq c_1\|\text{curl } \chi\| \quad \forall \chi \in J^1_\nu(\Omega),$$
$$\|\chi\|_{\mathscr{L}^2(\Omega)} \leqq c_2\|\text{curl } \chi\| \quad \forall \chi \in J^1_\tau(\Omega).$$

Then

$$|X_2| \leqq c_2(\|\text{curl } \varphi(T)\|\,\|\text{curl } \psi(T)\|+\|\text{curl } \varphi^0\|\,\|\text{curl } \psi^0\|)$$
$$\leqq (c_2/\sqrt{\varepsilon\mu})(\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2).$$

Define

$$(3.13)\qquad R(x_0) = \sup_{x\in\Omega}|m(x; x_0)|.$$

We have

$$|X_1| \leqq c_1R(x_0)(\|\text{curl } \varphi(T)\|\,\|\text{curl } \psi(T)\|+\|\text{curl } \varphi^0\|\,\|\text{curl } \psi^0\|)$$
$$\leqq (c_1/\sqrt{\varepsilon\mu})R(x_0)(\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2).$$

Therefore

$$(3.14)\qquad 2\varepsilon\mu|X_1+X_2| \leqq 2\sqrt{\varepsilon\mu}\,\max[c_1R(x_0), c_2](\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2).$$

Let us define

$$(3.15)\qquad T_0 = 2\sqrt{\varepsilon\mu}\,\max[c_1R(x_0), c_2].$$

From (3.12)-(3.15) we have

$$(3.16)\qquad (T-T_0)(\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2) \leqq \mu\int_\Sigma m\cdot\nu(\varepsilon\mu|\psi'|^2-|\text{curl } \psi|^2)\, d\Gamma\, dt.$$

We now introduce a *geometric assumption*: Γ is star-shaped with respect to some point $x_0\in\Omega$, i.e.,

$$(3.17)\qquad m\cdot\nu \geqq 0 \quad \text{on } \Gamma.$$

With (3.17), (3.16) simplifies to

$$(3.18)\qquad (T-T_0)(\varepsilon\|\text{curl } \varphi^0\|^2+\mu\|\text{curl } \psi^0\|^2) \leqq \varepsilon\mu^2\int_\Sigma m\cdot\nu|\psi'|^2\, d\Gamma\, dt.$$

Formula (3.18) has been obtained under assumption (3.3) on the data φ^0, ψ^0. Assume that $T > T_0$, and introduce the *norm*

$$(3.19)\qquad \|\{\varphi^0, \psi^0\}\|_{F_1} = \left[\int_\Sigma |\psi'|^2\, d\Gamma\, dt\right]^{1/2}.$$

Let F_1 be the completion of $J_\tau^*(\Omega)\times J_\nu^*(\Omega)$ with respect to the norm (3.19). It follows from standard trace theory that the imbedding $J_\tau^*(\Omega)\times J_\nu^*(\Omega)\to F_1$ is continuous. From (3.18) we have

$$F_1\subset J_\tau^1(\Omega)\times J_\nu^1(\Omega) \tag{3.20}$$

algebraically and topologically.

LEMMA 3.1. *Assume that* Γ *satisfies* (3.17) *and that* $T>T_0$, *where* T_0 *is defined in* (3.15). *Then for all* $\{\varphi^0,\psi^0\}\in F_1$,

$$(T-T_0)(\varepsilon\|\operatorname{curl}\varphi^0\|^2+\mu\|\operatorname{curl}\psi^0\|^2)\leqq\varepsilon\mu^2R(x_0)\int_\Sigma|\psi'|^2\,d\Gamma\,dt. \tag{3.21}$$

Remark 3.1. Another characterization of F_1 is

$$F_1=\{\{\varphi^0,\psi^0\}|\varphi^0\in J_\tau^1(\Omega),\ \psi^0\in J_\nu^1(\Omega),\ \psi|_\Sigma\in H^1(0,T;\mathscr{L}^2(\Gamma))\}.$$

If the star-shaped condition (3.17) is not satisfied, we obtain from (3.16) the estimate

$$(T-T_0)(\varepsilon\|\operatorname{curl}\varphi^0\|^2+\mu\|\operatorname{curl}\psi^0\|^2)\leqq\varepsilon\mu R(x_0)\int_\Sigma(\mu|\psi'|^2+\varepsilon|\varphi'|^2)\,d\Gamma\,dt. \tag{3.22}$$

Assume that $T>T_0$ and introduce the space F_2, the completion of $J_\tau^*(\Omega)\times J_\nu^*(\Omega)$ with respect to the norm

$$\|\{\varphi^0,\psi^0\}\|_{F_2}=\left[\int_\Sigma(|\psi'|^2+|\varphi'|^2)\,d\Gamma\,dt\right]^{1/2}.$$

We then have from (3.22)

$$F_2\subset J_\tau^1(\Omega)\times J_\nu^1(\Omega)$$

algebraically and topologically.

LEMMA 3.2. *Assume that* $T>T_0$, *defined in* (3.15). *Then for all* $\{\varphi^0,\psi^0\}\in F_2$,

$$(T-T_0)(\varepsilon\|\operatorname{curl}\varphi^0\|^2+\mu\|\operatorname{curl}\psi^0\|^2)\leqq\varepsilon\mu\max(\mu,\varepsilon)R(x_0)\int_\Sigma(|\psi'|^2+|\varphi'|^2)\,d\Gamma\,dt. \tag{3.23}$$

We shall now use (3.18), (3.22) to obtain additional energy estimates. Let us assume that the initial values satisfy

$$\varphi^0\in J_\tau^1(\Omega),\qquad \psi^0\in J_\nu^1(\Omega).$$

From (1.4) we have

$$\varepsilon\varphi(t)-\varepsilon\varphi^0-\operatorname{curl}\int_0^t\psi(s)\,ds=0,\qquad \mu\psi(t)-\mu\psi^0+\operatorname{curl}\int_0^t\varphi(s)\,ds=0\quad\text{in } Q. \tag{3.24}$$

We introduce the functions

$$\vartheta(t)=\int_0^t\psi(s)\,ds+\vartheta^0,\qquad \chi(t)=\int_0^t\varphi(s)\,ds+\chi^0, \tag{3.25}$$

where ϑ^0, χ^0 are chosen so that

$$\chi^0\in J_\tau^*(\Omega),\qquad \vartheta^0\in J_\nu^*(\Omega), \tag{3.26}$$

$$\operatorname{curl}\vartheta^0=\varepsilon\varphi^0,\qquad \operatorname{curl}\chi^0=-\mu\psi^0. \tag{3.27}$$

Let us assume for the moment that ϑ^0, χ^0 can be so determined. From (3.24)-(3.27) we have

$$\varepsilon\chi'-\operatorname{curl}\vartheta=0,\quad \mu\vartheta'+\operatorname{curl}\chi=0,\quad \operatorname{div}\chi=\operatorname{div}\vartheta=0\quad\text{in } Q, \tag{3.28}$$

$$\nu \times \chi = 0 \quad \text{on } \Sigma, \tag{3.29}$$

$$\chi(0) = \chi^0 \in J^*_\tau(\Omega), \qquad \vartheta(0) = \vartheta^0 \in J^*_\nu(\Omega). \tag{3.30}$$

We may therefore apply (3.18) (or (3.22)) to the solution of (3.28)–(3.30) and thereby obtain the following energy estimates:

$$(T - T_0)(\varepsilon \|\varphi^0\|^2 + \mu \|\psi^0\|^2) \leqq \mu \int_\Sigma m \cdot \nu |\psi|^2 \, d\Gamma \, dt \tag{3.31}$$

provided (3.17) holds; otherwise

$$(T - T_0)(\varepsilon \|\varphi^0\|^2 + \mu \|\psi^0\|^2) \leqq R(x_0) \int_\Sigma (\mu |\psi|^2 + \varepsilon |\varphi|^2) \, d\Gamma \, dt. \tag{3.32}$$

Assume that $T > T_0$, and introduce the norm

$$\|\{\varphi^0, \psi^0\}\|_{G_1} = \left[\int_\Sigma |\psi|^2 \, d\Gamma \, dt \right]^{1/2}$$

provided (3.17) is satisfied, and the norm

$$\|\{\varphi^0, \psi^0\}\|_{G_2} = \left[\int_\Sigma (|\psi|^2 + |\varphi|^2) \, d\Gamma \, dt \right]^{1/2}$$

in the general case. Define

$$G_1 = \text{completion of } J^1_\tau(\Omega) \times J^1_\nu(\Omega) \text{ with respect to } \| \cdot \|_{G_1},$$

$$G_2 = \text{completion of } J^1_\tau(\Omega) \times J^1_\nu(\Omega) \text{ with respect to } \| \cdot \|_{G_2}.$$

From (3.31), (3.32) we have

$$G_1 \subset J(\Omega) \times \hat{J}(\Omega) \tag{3.33}$$

algebraically and topologically whenever (3.17) is satisfied, and

$$G_2 \subset J(\Omega) \times \hat{J}(\Omega) \tag{3.34}$$

algebraically and topologically in the general case.

We have therefore proved (modulo (3.26), (3.27)) the following.

LEMMA 3.3. *Assume that* Γ *satisfies* (3.17) *and that* $T > T_0$, *where* T_0 *is defined in* (3.15). *Then for all* $\{\varphi^0, \psi^0\} \in G_1$,

$$(T - T_0)(\varepsilon \|\varphi^0\|^2 + \mu \|\psi^0\|^2) \leqq \mu R(x_0) \int_\Sigma |\psi|^2 \, d\Gamma \, dt.$$

LEMMA 3.4. *Assume that* $T > T_0$, *as defined in* (3.15). *Then for all* $\{\varphi^0, \psi^0\} \in G_2$,

$$(T - T_0)(\varepsilon \|\varphi^0\|^2 + \mu \|\psi^0\|^2) \leqq R(x_0) \max(\varepsilon, \mu) \int_\Sigma (|\psi|^2 + |\varphi|^2) \, d\Gamma \, dt.$$

Remark 3.2. From the estimates above we obtain the following *uniqueness results* for Maxwell's equations:

(i) Let φ, ψ be a solution to the Maxwell system (1.4), and assume that $T > T_0$, defined in (3.15). Then

$$\psi = \varphi = 0 \quad \text{on } \Sigma \Rightarrow \{\varphi, \psi\} = 0 \quad \text{in } \Omega \times (0, T).$$

(ii) If $T > T_0$ and (3.17) holds, then

$$\psi = \nu \times \varphi = 0 \quad \text{on } \Sigma \Rightarrow \{\varphi, \psi\} = 0 \quad \text{in } \Omega \times (0, T).$$

Proof of (3.26), (3.27). Since the map $\varphi \to \operatorname{curl} \varphi$ is a homeomorphism from $J^2_\nu(\Omega)$ (respectively, $J^2_\tau(\Omega)$) onto $J^1_\tau(\Omega)$ (respectively, $J^1_\nu(\Omega)$), there are functions $\tilde{\varphi}^0 \in J^2_\nu(\Omega)$,

$\tilde{\varphi}^0 \in J_\tau^2(\Omega)$ such that

$$\varphi^0 = \operatorname{curl} \tilde{\varphi}^0, \qquad \psi^0 = \operatorname{curl} \tilde{\psi}^0.$$

We define

$$\vartheta^0 = \varepsilon\tilde{\varphi}^0 + \varepsilon\nabla f, \qquad \chi^0 = -\mu\tilde{\psi}^0 + \mu\nabla g,$$

where f, g are chosen according to

$$\nabla^2 f = -\operatorname{div} \tilde{\varphi}^0 \quad \text{in } \Omega, \qquad \frac{\partial f}{\partial \nu} = -\nu \cdot \tilde{\varphi}^0 \quad \text{on } \Gamma,$$

$$\nabla^2 g = \operatorname{div} \tilde{\psi}^0 \quad \text{in } \Omega, \qquad g = 0 \quad \text{on } \Gamma.$$

Then (3.27) is satisfied and

$$\nu \cdot \vartheta^0 = 0, \qquad \nu \times \chi^0 = \nu \times \nabla g = 0 \quad \text{on } \Gamma,$$

the latter equality being a consequence of the fact that $\nu \times \nabla$ is a vector of *tangential* first-order differential operators in Γ. Thus χ^0, ϑ^0 satisfy (3.26).

4. Exact controllability. In this section we will use the energy estimates of § 3 in conjunction with HUM to prove exact controllability to rest of solutions of (1.1)-(1.3). Several different results will be presented, depending on the geometry of Γ and the regularity of the initial data. We note that proving controllability to rest at time T is equivalent to proving controllability to an arbitrary state (in a suitable function space) at time T since, for the *homogeneous* problem, the map $\{E^0, H^0\} \to \{E(T), H(T)\}$ is an isomorphism in the appropriate spaces.

4.1. Exact controllability in $J(\Omega) \times \hat{J}(\Omega)$ with $\mathscr{L}^2(\Sigma)$ boundary controls, under a geometric assumption on Γ. In this section we suppose that Γ satisfies (3.17). Assume that $T > T_0$. Let $\{\varphi^0, \psi^0\} \in G_1$, and let $\{\varphi, \psi\}$ be the solution to (1.4)-(1.6). Then $\psi|_\Sigma \in \mathscr{L}^2(\Sigma)$. Consider the following problem (where $\hat{\varepsilon} = 1/\varepsilon$, $\hat{\mu} = 1/\mu$):

$$\frac{\partial E}{\partial t} - \hat{\varepsilon} \operatorname{curl} H = 0, \quad \frac{\partial H}{\partial t} + \hat{\mu} \operatorname{curl} E = 0, \quad \operatorname{div} E = \operatorname{div} H = 0 \quad \text{in } Q, \tag{4.1}$$

$$E(T) = H(T) = 0 \quad \text{in } \Omega, \tag{4.2}$$

$$\nu \times H = \psi|_\Sigma \quad \text{on } \Sigma. \tag{4.3}$$

We will see below that (4.1)-(4.3) has a unique solution (in a weak sense). Let us form the expression

$$0 = \int_Q \left[\left(\frac{\partial E}{\partial t} - \hat{\varepsilon} \operatorname{curl} H \right) \cdot \psi - \left(\frac{\partial H}{\partial t} + \hat{\mu} \operatorname{curl} E \right) \cdot \varphi \right] dx\, dt. \tag{4.4}$$

Proceeding formally (everything will be justified in Remark 4.2 below), after an integration by parts we obtain from (4.4)

$$\begin{aligned} &-(E(0), \psi^0) + (H(0), \varphi^0) - \int_Q \left(E \cdot \frac{\partial \psi}{\partial t} - H \cdot \frac{\partial \varphi}{\partial t} \right) dx\, dt \\ &- \int_Q (\hat{\varepsilon}\psi \cdot \operatorname{curl} H + \hat{\mu}\varphi \cdot \operatorname{curl} E)\, dx\, dt = 0. \end{aligned} \tag{4.5}$$

The second integral in (4.5) may be written as

$$\int_Q (\hat{\varepsilon}H \cdot \operatorname{curl} \psi + \hat{\mu}E \cdot \operatorname{curl} \varphi)\, dx\, dt - \int_\Sigma [\nu \cdot (H \times \psi) + \nu \cdot (E \times \varphi)]\, d\Gamma\, dt$$

$$= \int_Q (\hat{\varepsilon}H \cdot \operatorname{curl} \psi + \hat{\mu}E \cdot \operatorname{curl} \varphi)\, dx\, dt - \int_\Sigma |\psi|^2\, d\Gamma\, dt.$$

Therefore (4.5) reduces to

$$(E(0), \psi^0) - (H(0), \varphi^0) = \int_\Sigma |\psi|^2\, d\Gamma\, dt,$$

which may be written as

$$\langle \{-H(0), E(0)\}, \{\varphi^0, \psi^0\} \rangle = \int_\Sigma |\psi|^2\, d\Gamma\, dt. \tag{4.6}$$

Let us define a linear mapping Λ by

$$\Lambda\{\varphi^0, \psi^0\} = \{-H(0), E(0)\}. \tag{4.7}$$

In terms of Λ, (4.6) becomes

$$\langle \Lambda\{\varphi^0, \psi^0\}, \{\varphi^0, \psi^0\} \rangle = \|\{\varphi^0, \psi^0\}\|^2_{G_1}, \tag{4.8}$$

and therefore Λ *is an isomorphism from* G_1 *onto* G_1' (G_1' = dual of G_1 with respect to $J(\Omega) \times \hat{J}(\Omega)$). Consequently, if $\{-H^0, E^0\} \in G_1'$ and if we choose $\{\varphi^0, \psi^0\} = \Lambda^{-1}\{-H^0, E^0\}$, then the unique solution of (4.1), (4.3) with initial data

$$E(0) = E^0, \qquad H(0) = H^0 \quad \text{in } \Omega$$

will (by construction) satisfy (4.2). Since $G_1' \supset J(\Omega) \times \hat{J}(\Omega)$, we have Theorem 4.1.

THEOREM 4.1. *Assume that* Γ *satisfies* (3.17), *that* $\{-H^0, E^0\} \in J(\Omega) \times \hat{J}(\Omega)$, *and that* $T > T_0$ *as defined in* (3.15). *Then the control* $J = -\psi|_\Sigma \in \mathscr{L}^2(\Sigma)$ *drives the system* (1.1)-(1.3) *to rest at time* T.

Remark 4.1. For initial data $\{-H^0, E^0\} \in J(\Omega) \times \hat{J}(\Omega)$, the control $J = -\psi|_\Sigma$ *minimizes* the norm $\int_\Sigma |J|^2\, d\Gamma\, dt$ among all controls $J \in \mathscr{L}^2(\Sigma)$ which drive the system (1.1)-(1.3) to rest at time T.

Remark 4.2. We still need to make precise the sense in which (4.1)-(4.3) (and (1.1)-(1.3)) are to be understood. This is done using the transposition method. In fact, let us consider the forward problem (1.1)-(1.3) with $J \in \mathscr{L}^2(\Sigma)$ and initial data satisfying $\{H^0, -E^0\} \in \mathscr{H}'$, where $\mathscr{H} = J_\tau^1(\Omega) \times J_\nu^1(\Omega)$ and $\mathscr{H}'$ is the dual of $\mathscr{H}$ with respect to $J(\Omega) \times \hat{J}(\Omega)$. Let $\{\varphi^0, \psi^0\} \in \mathscr{H}$ and $\{\varphi, \psi\}$ be the solution to (1.4)-(1.6). Consider the expression

$$0 = \int_0^t \int_\Omega \left[\left(\frac{\partial E}{\partial t} - \hat{\varepsilon} \operatorname{curl} H \right) \cdot \psi - \left(\frac{\partial H}{\partial t} + \hat{\mu} \operatorname{curl} E \right) \cdot \varphi \right] dx\, ds.$$

If we apply integration by parts in t and Green's formula in x we are led to

$$(H(t), \varphi(t)) - (E(t), \psi(t)) = (H^0, \psi^0) - (E^0, \psi^0) - \int_0^t \int_\Gamma J \cdot \psi\, d\Gamma\, ds. \tag{4.9}$$

We rewrite (4.9) as

$$\langle \{H(t), -E(t)\}, \{\varphi(t), \psi(t)\} \rangle = \langle \{H^0, -E^0\}, \{\varphi^0, \psi^0\} \rangle - \int_0^t \int_\Gamma J \cdot \psi\, d\Gamma\, ds. \tag{4.10}$$

Equation (4.10) is the *definition* of the solution of (1.1)-(1.3). In fact, we have

$$\begin{aligned}\left|\langle\{H^0,-E^0\},\{\varphi^0,\psi^0\}\rangle-\int_0^t\int_\Gamma J\cdot\psi\,d\Gamma\,ds\right| &\leqq \|\{H^0,-E^0\}\|_{\mathscr{K}'}\|\{\varphi^0,\psi^0\}\|_{\mathscr{K}}+\|J\|_{\mathscr{L}^2(\Sigma)}\|\psi\|_{\mathscr{L}^2(\Sigma)}\\ &\leqq \|\{H^0,-E^0\}\|_{\mathscr{K}'}\|\{\psi^0,\psi^0\}\|_{\mathscr{K}}+c_1\|J\|_{\mathscr{L}^2(\Sigma)}\|\operatorname{curl}\psi\|_{\mathscr{L}^2(\Sigma)}\\ &\leqq [\|\{H^0,-E^0\}\|_{\mathscr{K}'}+c_1\|J\|_{\mathscr{L}^2(\Sigma)}]\|\{\varphi^0,\tilde{\psi^0}\}\|_{\mathscr{K}}.\end{aligned}\tag{4.11}$$

Since the map $\{\varphi^0,\psi^0\}\to\{\varphi(t),\psi(t)\}\colon\mathscr{K}\to\mathscr{K}$ is an *isomorphism*, it follows from (4.11) that there is a unique pair $\{H,-E\}\in L^\infty(0,T;\mathscr{K}')$ that satisfies (4.10). (In fact, $\{H,-E\}\in C([0,T];\mathscr{K}')$.) This pair is, by definition, the solution to (1.1)-(1.3). (The fact that $T>T_0$ plays no role in the existence theory so far.)

Now suppose that $T>T_0$ and that $\{H^0,-E^0\}\in G_1'$ (recall that $G_1'\subset\mathscr{K}'$). From (4.10) with $t=T$ we obtain

$$|\langle\{H(T),E(T)\},\{\varphi(T),-\psi(T)\}\rangle|\leqq[\|\{H^0,-E^0\}\|_{G_1'}+\|J\|_{\mathscr{L}^2(\Sigma)}]\|\{\varphi^0,\psi^0\}\|_{G_1}.$$

From the definition of G_1 it can be seen that $\{\varphi^0,\psi^0\}\in G_1$ if, and only if, $\{\varphi(T),-\psi(T)\}\in G_1$; further, the map $\{\varphi^0,\psi^0\}\to\{\varphi(T),-\psi(T)\}\colon G_1\to G_1$ is an isometry. Thus, $\{H^0,-E^0\}\in G_1'$ implies that $\{H(T),E(T)\}\in G_1'$.

For the problem (4.1)-(4.3), it follows from the preceding discussion that there exists a unique solution $\{H,-E\}\in C([0,T];\mathscr{K}')$, and $\{H(0),\pm E(0)\}\in G_1'$. In particular, (4.6) holds *by definition* of the solution.

4.2. Exact controllability in $(J_\tau^1(\Omega))'\times(J_\nu^1(\Omega))'$ with $(H^1(0,T;\mathscr{L}^2(\Gamma)))'$ boundary controls, under a geometric assumption on Γ. We may extend the space of initial data that can be exactly controlled to rest beyond the space G_1', but at the expense of working with controls weaker than the $\mathscr{L}^2(\Sigma)$ controls of Theorem 4.1.

Let $\{\varphi^0,\psi^0\}\in J_\tau^*(\Omega)\times J_\nu^*(\Omega)$, assume that $T>T_0$ and consider the problem consisting of the Maxwell system (4.1), the terminal data (4.2) and the boundary condition

$$\nu\times H=-\psi''|_\Sigma\quad\text{on }\Sigma.\tag{4.12}$$

Formally applying Green's formula as in § 4.1, we obtain (analogous to (4.6))

$$\langle\{-H(0),E(0)\},\{\varphi^0,\psi^0\}\rangle=-\int_\Sigma\psi\cdot\psi''\,d\Gamma\,dt.\tag{4.13}$$

In (4.12), (4.13) we *define* $\psi''|_\Sigma\in(H^1(0,T;\mathscr{L}^2(\Gamma)))'$ by the duality

$$\langle\psi'',\chi\rangle=-\int_\Sigma\psi'\cdot\chi'\,d\Gamma\,dt\quad\forall\chi\in H^1(0,T;\mathscr{L}^2(\Gamma)).\tag{4.14}$$

Equation (4.14) has a meaning whenever $\{\varphi^0,\psi^0\}\in F_1$. Note that $\psi''|_\Sigma$ is *not* the distributional derivative $d/dt(\psi')\in H^{-1}(0,T;\mathscr{L}^2(\Gamma))$.

With the mapping Λ defined in (4.7) we obtain from (4.13), (4.14)

$$\langle\Lambda\{\varphi^0,\psi^0\},\{\varphi^0,\psi^0\}\rangle=\int_\Sigma|\psi'|^2\,d\Gamma\,dt=\|\{\varphi^0,\psi^0\}\|_{F_1}^2.\tag{4.15}$$

Consequently, Λ is an isomorphism from F_1 into F_1'. Since $F_1'\supset(J_\tau^1(\Omega))'\times(J_\nu^1(\Omega))'$ we have Theorem 4.2.

THEOREM 4.2. *Assume that* Γ *satisfies* (3.17), *that* $\{-H^0, E^0\} \in (J_\tau^1(\Omega))' \times (J_\nu^1(\Omega))'$, *and that* $T > T_0$ *as defined in* (3.15). *Then the control* $J = -\psi'|_\Sigma \in (H^1(0, T; \mathscr{L}^2(\Gamma)))'$ *drives the system* (1.1)-(1.3) *to rest at time* T.

Remark 4.3. For initial data $\{H^0, -E^0\}$ and boundary data J satisfying

$$\{H^0, -E^0\} \in (J_\tau^*(\Omega))' \times (J_\nu^*(\Omega))' \doteq \mathscr{Y}', \qquad J \in (H^1(0, T; \mathscr{L}^2(\Gamma)))',$$

the problem (1.1)-(1.3) may once again be solved by transposition, using (4.10) as the definition of the solution. We get $\{H, -E\} \in C([0, T]; \mathscr{Y}')$ with

$$\|\{H, -E\}\|_{L^\infty(0, T; \mathscr{Y}')} \leqq \|\{H^0, -E^0\}\|_{\mathscr{Y}'} + c_1 \|J\|_{(H^1(0, T; \mathscr{L}^2(\Gamma)))'}.$$

If, in particular, $\{H^0, -E^0\} \in F_1'$, then $\{H(T), E(T)\} \in F_1'$.

4.3. Exact controllability without geometric restrictions on Γ. If the geometric restriction (3.17) on Γ is lifted, we must work with the spaces F_2 and G_2 rather than F_1 and G_1 when applying HUM. As usual, the object is to choose the boundary control so that Λ is an isomorphism of F_2 (respectively, G_2) onto F_2' (respectively, G_2'). We then obtain exact controllability in the space F_2' (respectively, G_2'). It is shown below that exact controllability in the space F_2' can be established in this manner. However, we are unable to prove exact controllability in the space G_2'.

One possible explanation for the lack of (provable) exact controllability in G_2' is the following. The topology of G_2 is weaker than that of F_2, and therefore the opposite is true of their duals. Consequently, exact controllability in G_2' means controllability of states that are smoother than those in F_2', by means of controls more regular than those used to control states in F_2'. Because no restrictions of a geometric nature are assumed regarding Γ, it is not implausible that there are smooth initial states that *cannot* be steered to rest using such regular controls, i.e., exact controllability in G_2' may not be possible without geometric restrictions on Γ.

To establish exact controllability in F_2', we consider as usual for $T > T_0$ the Maxwell system (4.1) with terminal data (4.2) and boundary data

$$\nu \times H = -J \quad \text{on } \Sigma. \tag{4.16}$$

We consider also the homogeneous problem (1.4)-(1.6) with initial data $\{\varphi^0, \psi^0\} \in J_\tau^*(\Omega) \times J_\nu^*(\Omega)$. The solution of (4.1), (4.2), and (4.16) is defined by transposition. In particular, the pair $\{-H(0), E(0)\}$ satisfies (by definition!)

$$\langle \{-H(0), E(0)\}, \{\varphi^0, \psi^0\} \rangle = -\int_\Sigma \psi \cdot J \, d\Gamma dt. \tag{4.17}$$

The object is to choose the control J such that

$$-\int_\Sigma \psi \cdot J \, d\Gamma \, dt = \|\{\varphi^0, \psi^0\}\|_{F_2}^2 = \int_\Sigma (|\psi'|^2 + |\operatorname{curl} \psi|^2) \, d\Gamma \, dt. \tag{4.18}$$

With (4.17), (4.18), we have that Λ is an isomorphism of F_2 onto F_2', which proves exact controllability in that space.

Now (4.18) will be satisfied if

$$J = -\psi'|_\Sigma + \chi,$$

where χ satisfies

$$-\int_\Sigma \chi \cdot \psi \, d\Gamma \, dt = \int_\Sigma |\operatorname{curl} \psi|^2 \, d\Gamma \, dt.$$

To determine χ, we first note that for sufficiently smooth functions $\hat{\psi}: \Omega \to \mathbb{R}^3$ we have

$$\frac{\partial \hat{\psi}_i}{\partial x_k} = \nu_k \frac{\partial \hat{\psi}_i}{\partial \nu} + \sigma_k \hat{\psi}_i \quad \text{on } \Gamma,$$

where σ_k is a tangential operator in Γ of order 1. Therefore

$$\operatorname{curl} \hat{\psi} = \nu \times \frac{\partial \hat{\psi}}{\partial \nu} + \sigma \times \hat{\psi} \quad \text{on } \Gamma.$$

The operator $\psi \to \sigma \times \psi$ is a formally self-adjoint operator on Γ: for all sufficiently smooth functions $\hat{\varphi}$, $\hat{\psi}$ defined on Γ,

$$\int_\Gamma \hat{\varphi} \cdot (\sigma \times \hat{\psi})\, d\Gamma = \int_\Gamma (\sigma \times \hat{\varphi}) \cdot \hat{\psi}\, d\Gamma. \tag{4.19}$$

Since $\{\varphi^0, \psi^0\} \in J_\tau^*(\Omega) \times J_\nu^*(\Omega)$, $(\operatorname{curl} \psi)|_\Gamma$ and $\varphi'|_\Gamma$ are defined in the sense of traces and therefore $\operatorname{curl} \psi = \varepsilon \varphi'$ on Σ. As a result,

$$|\operatorname{curl} \psi|^2 = \varepsilon \varphi' \cdot \operatorname{curl} \psi = \varepsilon \varphi' \cdot \left(\nu \times \frac{\partial \psi}{\partial \nu} \right) + \varepsilon \varphi' \cdot (\sigma \times \psi) = \varepsilon \varphi' \cdot (\sigma \times \psi) \quad \text{on } \Sigma,$$

since $\nu \times \varphi = 0$ on Σ. Consequently,

$$\int_\Sigma |\operatorname{curl} \psi|^2\, d\Gamma\, dt = \varepsilon \int_\Sigma \varphi' \cdot (\sigma \times \psi)\, d\Gamma\, dt = \varepsilon \int_\Sigma \psi \cdot (\sigma \times \varphi')\, d\Gamma\, dt.$$

Therefore, if we choose the control J according to

$$J = \psi''|_\Sigma - \varepsilon \sigma \times \varphi',$$

then (4.18) will hold.

Remark 4.4. It follows from (4.19) that $\sigma \times \varphi' \in L^2(0, T; \mathscr{H}^{-1}(\Gamma))$.

THEOREM 4.3. *Assume that* $\{-H^0, E^0\} \in (J_\tau^1(\Omega))' \times (J_\nu^1(\Omega))'$ *and that* $T > T_0$ *as defined in* (3.15). *Then the control*

$$J = \psi''|_\Sigma - \varepsilon \sigma \times \varphi' \in (H^1(0, T; \mathscr{L}^2(\Gamma)))' \oplus L^2(0, T; \mathscr{H}^{-1}(\Gamma)) \tag{4.20}$$

drives the system (1.1)-(1.3) *to rest at time T.*

REFERENCES

[1] K. O. FRIEDRICHS, *Mathematical methods of electromagnetic theory*, Courant Institute of Mathematical Sciences, New York University, New York, 1974.

[2] K. A. KIME, *Boundary controllability of Maxwell's equations in a spherical region*, SIAM J. Control Optim., submitted.

[3] O. A. LADYZHENSKAYA AND V. A. SOLONIKOV, *The linearization principle and invariant manifolds for problems of magnetohydrodynamics*, J. Soviet Math., 8 (1977), pp. 384-422.

[4] J. L. LIONS, *Contrôlabilité exacte des systèmes distribués*, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 471-475.

[5] ———, *Exact controllability, stabilization and perturbations for distributed systems*, John Von Neumann Lecture, Boston, MA, July, 1986; SIAM Rev., 30 (1988), pp. 1-68.

[6] D. L. RUSSELL, *The Dirichlet-Neumann boundary control problem associated with Maxwell's equations in a cylindrical region*, SIAM J. Control Optim., 24 (1986), pp. 199-229.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 389-406, March 1989

008

ROBUST STABILITY AND PERFORMANCE VIA FIXED-ORDER DYNAMIC COMPENSATION*

DENNIS S. BERNSTEIN†

Abstract. Two robust control-design problems are considered. The Robust Stabilization Problem involves deterministically modeled, bounded but unknown, time-varying parameter variations, while the Robust Performance Problem includes, in addition, a quadratic performance criterion averaged over stochastic disturbances and maximized over the admissible parameter variations. For both problems the design goal is a fixed-order (i.e., reduced- or full-order) dynamic (strictly proper) feedback compensator. A sufficient condition for solving the Robust Stabilization Problem is given by means of a quadratic Lyapunov function parameterized by the compensator gains. For the Robust Performance Problem the Lyapunov function provides an upper bound for the closed-loop performance. This leads to consideration of the Auxiliary Minimization Problem: Minimize the performance bound over the class of fixed-order controllers subject to the Lyapunov-function constraint. Necessary conditions for optimality in the auxiliary problem thus serve as sufficient conditions for robust stability and performance in the original problem. Two particular bounds are considered for constructing the quadratic Lyapunov function. The first corresponds to a right shift/multiplicative white noise model, while the second was suggested by recent work of Petersen and Hollot. The main result is an extended version of the optimal projection equations for fixed-order dynamic compensation whose solutions are guaranteed to provide both robust stability and robust performance.

Key words. robust control, stability, performance, dynamic compensation, Lyapunov bounds

AMS(MOS) subject classification. 93

1. Introduction.

Although considerable effort has been devoted to frequency-domain robust-control design methods [1]-[10], there remain open questions concerning stability with respect to real-valued, structured plant parameter variations [11]-[13]. Specifically, it is shown in [11]-[13] that classical gain and phase margin specifications can be satisfied, while sensitivity to structured plant parameter variations can be arbitrarily large. From a time-domain point of view, the parametric robustness problem has been widely studied using Lyapunov's second method as the principal technique [14]-[28].

In this paper we develop an approach to control design that provides sufficient conditions for robust stability and performance over a prescribed range of time-varying structured plant parameter variations by means of a feedback law in the form of a fixed-order (i.e., reduced- or full-order) dynamic (strictly proper) compensator. The approach is based upon the merging of two techniques, namely, the guaranteed cost control approach to robust performance [14], [17] and the optimal projection approach to quadratically optimal fixed-order dynamic compensation [29], [30]. One of our goals is to obtain robust output-feedback compensators rather than full-state-feedback controllers. Also, since we wish to account for real-time computational burden in implementating the controller, we impose a constraint on the dimension (i.e., order) of the dynamic compensator. This approach thus generalizes standard LQG theory, which yields full-order output-feedback controllers for systems without parameter uncertainty. We note that our approach is constructive in the sense that, upon satisfaction of the sufficient conditions, the feedback gains required for implementing the robust feedback controller are explicitly synthesized. Existential issues are also addressed

* Received by the editors June 16, 1986; accepted for publication (in revised form) May 10, 1988. This research was supported in part by Air Force Office of Scientific Research contract F49620-86-C-0002.

† Harris Corporation, Melbourne, Florida 32901.

herein, although to a lesser extent. For further background see [29], [30]. For extensions to nonstrictly proper controllers see [31], and for extensions to H_∞ control see [32].

To explain the rationale behind the development we briefly describe the main elements of the approach. The following discussion is intended to be descriptive; precise conditions appear in the main body of the paper.

1.1. Robust Stability Problem. For a nominal linear time-invariant (A, B, C) system we consider deterministically modeled bounded but otherwise unknown Lebesgue measurable time-varying parameter variations of the form

$$A+\sum_{i=1}^{p}\hat{\sigma}_i(t)A_i, \quad B+\sum_{i=1}^{p}\hat{\sigma}_i(t)B_i, \quad C+\sum_{i=1}^{p}\hat{\sigma}_i(t)C_i. \tag{1.1}$$

The nominal matrices A, B, C and the perturbation matrices A_i, B_i, C_i denoting the structure of the parametric uncertainty are assumed known, while the time-varying uncertain parameters $\hat{\sigma}_i(t)$ are assumed only to satisfy the bounds

$$|\hat{\sigma}_i(t)| \leqq \delta_i, \quad i=1,\cdots,p, \quad t\in[0,\infty). \tag{1.2}$$

The form of (1.1) permits an arbitrary number of uncertain parameters with arbitrary linear structure. Although we do not require matching conditions as in [21], the linear structure of (1.1) is more restrictive than the functional form $A(q(t))$ used in [21]. It is this structure that we exploit to obtain sufficiency conditions. Note also that the representation (1.1) is independent of state space basis, since replacing A by SAS^{-1} corresponds to replacing A_i by SA_iS^{-1}. As will be seen, our robustness bounds and optimality conditions are also basis independent. Also, scaling techniques [6], [7] will not play a role here. Finally, we note that because of the time-varying nature of the uncertain perturbations (1.1) it is virtually impossible to determine the *actual* stability region of a given design by means of empirical methods.

1.2. Quadratic Lyapunov function. As a sufficient condition for characterizing solutions of the Robust Stability Problem we consider a closed-loop quadratic Lyapunov function $V(\tilde{x})=\tilde{x}^T\mathcal{P}\tilde{x}$, where the matrix $\mathcal{P}$ satisfies

$$0=\tilde{A}^T\mathcal{P}+\mathcal{P}\tilde{A}+\Omega(\mathcal{P}, B_c, C_c) \tag{1.3}$$

and the function Ω is a bound satisfying

$$\sum_{i=1}^{p}\sigma_i(\tilde{A}_i^T\mathcal{P}+\mathcal{P}\tilde{A}_i)<\Omega(\mathcal{P}, B_c, C_c) \tag{1.4}$$

over the parameter range

$$|\sigma_i| \leqq \delta_i, \qquad i=1,\cdots,p. \tag{1.5}$$

Note that the constant σ_i in (1.4) and (1.5) plays the role of $\hat{\sigma}_i(t)$, i.e., t is "frozen" in (1.4) and (1.5). In (1.3) and (1.4), $\tilde{A}$ and $\tilde{A}_i$ denote the closed-loop dynamics and closed-loop parameter-uncertainty matrices given by

$$\tilde{A}=\begin{bmatrix} A & BC_c \\ B_cC & A_c \end{bmatrix}, \qquad \tilde{A}_i=\begin{bmatrix} A_i & B_iC_c \\ B_cC_i & 0 \end{bmatrix}. \tag{1.6}$$

Since $\tilde{A}_i$ is independent of A_c, Ω depends only on B_c and C_c. As discussed later in this section, (1.4) is automatically satisfied by construction of the function Ω. Furthermore, the existence of a solution $\mathcal{P}$ to (1.3) need not be verified directly, but rather is a result of numerically solving the optimality conditions discussed below.

1.3. Robust Performance Problem. In addition to the *deterministic* parameter uncertainty model (1.1), (1.2), the Robust Performance Problem includes *stochastic* plant disturbances and measurement noise with performance measured by means of the quadratic functional

$$\tilde{J}(t) = x^T(t)R_1x(t) + 2x^T(t)R_{12}u(t) + u^T(t)R_2u(t). \tag{1.7}$$

To obtain a steady-state design problem we (1) average $\tilde{J}(t)$ over the disturbance and measurement noise statistics; (2) pass to the steady-state limit; and (3) maximize over the class of parameter uncertainties. Hence the performance of a given controller (A_c, B_c, C_c) is given by

$$J(A_c, B_c, C_c) = \sup_{\tilde{\sigma}(\cdot)} \limsup_{t\to\infty} \mathbb{E}[\tilde{J}(t)]. \tag{1.8}$$

The use of "lim sup" is a technicality that accounts for cases in which the steady-state limit may not exist. Note that although (1.8) is an averaging criterion over the disturbances as in LQG theory, it is also a worst-case measure over the uncertain parameters. Thus (1.8) is a *hybrid* criterion in the sense that is *stochastic* in the disturbance space (i.e., external uncertainties) and *deterministic* in the parameter space (i.e., internal uncertainties). By "internal uncertainties" we have in mind quantities such as mass, damping, or stiffness; by "external uncertainties" we are referring to phenomena such as turbulent flow for which only power spectrum statistics may be available. No claim is made, however, with regard to the universal validity of such a mathematical uncertainty model. In particular applications, uncertainty models that are either wholly deterministic or wholly stochastic may be more appropriate. In general, our setting appears to be consistent with the available literature (see [1]-[28]).

1.4. Performance bound. To obtain a tractable design problem, we use the matrix $\mathcal{P}$ to bound the performance of each controller solving the Robust Stability Problem. Specifically, by assuming in addition to (1.4) that

$$\sum_{i=1}^{p} \sigma_i(\tilde{A}_i^T\mathcal{P} + \mathcal{P}\tilde{A}_i) + \tilde{R} \leqq \Omega(\mathcal{P}, B_c, C_c), \tag{1.9}$$

it follows that

$$J(A_c, B_c, C_c) \leqq \operatorname{tr} \mathcal{P}\tilde{V}. \tag{1.10}$$

In (1.9) and (1.10) $\tilde{R}$ and $\tilde{V}$ denote closed-loop weighting and disturbance intensity matrices. The idea of bounding the performance by means of a Lyapunov function is the basis for guaranteed cost control [14], [17].[1]

1.5. Construction of the Lyapunov function. So far the Lyapunov function has only been abstractly characterized by means of (1.3) and (1.4). To obtain a useful design theory Ω is now given a concrete form. Specifically, to satisfy (1.9) it is assumed that

$$\Omega(\mathcal{P}, B_c, C_c) = \sum_{i=1}^{p} \Lambda_i(\mathcal{P}, B_c, C_c) + \tilde{R}, \tag{1.11}$$

where, for each i, the Λ_i are chosen such that

$$\sigma_i(\tilde{A}_i^T\mathcal{P} + \mathcal{P}\tilde{A}_i) \leqq \Lambda_i(\mathcal{P}, B_c, C_c), \qquad |\sigma_i| \leqq \delta_i. \tag{1.12}$$

[1] It is also interesting to note that in Hamilton-Jacobi-Bellman sufficiency theory the performance functional is expressed in terms of a value function that also serves as a Lyapunov function for the closed-loop system. These connections will be explored in a future paper.

Note that (1.12) implies that (1.4) holds with Ω given by (1.11). Since $\tilde{A}_i$ depends on B_c and C_c, the bound Λ_i will be constructed to be *gain-invariant*, that is, so that (1.12) holds for *all* B_c and C_c. Thus no difficulty will arise from the fact that the controller gains are yet to be determined by optimality considerations.

It should be noted that the bounding in (1.12) is defined in the sense of the cone of nonnegative definite matrices. Since this is only a partial ordering and not a total ordering, a least upper bound (i.e., a "sharpest" bound) does not exist in general and the conservatism of the inequality in (1.12) cannot be quantified by a scalar measure. Hence, Λ_i satisfying (1.12) is not necessarily unique and two particular choices of Λ_i are developed in this paper. Since we shall utilize first-order necessary conditions for optimality, we confine our consideration to bounds that are differentiable. The first choice of Λ_i satisfying (1.12) is given by the linear (in $\mathscr{P}$) function

$$\Lambda_i(\mathscr{P}, B_c, C_c) = \delta_i(\alpha_i \mathscr{P} + \alpha_i^{-1} \tilde{A}_i^T \mathscr{P} \tilde{A}_i), \tag{1.13}$$

where α_i is an arbitrary positive number. As shown in [33], the bound (1.13) can be viewed as arising from a stochastic optimal control problem with exponentially weighted cost and state-, control- and measurement-dependent white noise. The stochastic multiplicative white noise model serves only as an *interpretation*, however, and need not be viewed as having physical significance. A similar bound is used in [28].

The second choice for Λ_i satisfying (1.12) is given by the quadratic (in $\mathscr{P}$) function

$$\Lambda_i(\mathscr{P}, B_c, C_c) = \delta_i(\tilde{E}_i^T \tilde{E}_i + \mathscr{P} \tilde{D}_i \tilde{D}_i^T \mathscr{P}), \tag{1.14}$$

where $\tilde{D}_i$, $\tilde{E}_i$ denote an arbitrary factorization of $\tilde{A}_i$ of the form

$$\tilde{A}_i = \tilde{D}_i \tilde{E}_i. \tag{1.15}$$

The bound (1.14) was used in [26] for full-state feedback with rank 1 uncertainties. Note that using congruence transformations shows that both bounds (1.13) and (1.14) are basis independent; that is, replacing $\tilde{A}_i$ by $\tilde{S}\tilde{A}_i\tilde{S}^{-1}$ leads to replacing $\mathscr{P}$ by $\tilde{S}^{-T}\mathscr{P}\tilde{S}^{-1}$.

1.6. Auxiliary Minimization Problem. The next step in our development for robust performance is the following. Inasmuch as the performance of a robustly stabilizing controller is bounded via (1.10) over the given range of parameter variations, it is desirable to minimize the upper bound

$$\mathscr{J}(\mathscr{P}, A_c, B_c, C_c) \triangleq \operatorname{tr} \mathscr{P} \tilde{V} \tag{1.16}$$

subject to the constraint (1.3). This is referred to as the Auxiliary Minimization Problem. For a given choice (1.13) or (1.14) of Λ_i for each i, a solution of the Auxiliary Minimization Problem provides a controller whose steady-state performance is guaranteed to remain below the bound (1.16) over the range of parameter variations, hence guaranteeing robust performance. Since the Auxiliary Minimization Problem is a smooth mathematical programming problem, a minimum always exists on compact sets. To actually characterize extremals of the Auxiliary Minimization Problem we proceed by deriving first-order necessary conditions. Because these necessary conditions are derived for the Auxiliary Minimization Problem, they effectively serve as sufficient conditions for robustness in the original problem.

It should be noted that the guaranteed cost control approach developed in [14] does not permit this line of development since Λ_i is given by

$$\Lambda_i(\mathscr{P}, B_c, C_c) = \delta_i |\tilde{A}_i^T \mathscr{P} + \mathscr{P} \tilde{A}_i|, \tag{1.17}$$

where $|\cdot|$ denotes the matrix obtained by replacing each eigenvalue by its absolute value. Since this bound is not differentiable with respect to the controller gains, first-order necessary conditions cannot be used.

1.7. The optimality conditions: full-order case. For the full-order case, i.e., when the order of the controller is equal to the order of the plant, the first-order necessary conditions can be derived in a form that is a direct generalization of the pair of separated Riccati equations of LQG theory. Specifically, the necessary conditions comprise a coupled system of four algebraic matrix equations including a pair of modified Riccati equations and a pair of Lyapunov equations. For plant models involving multiplicative white noise these equations have been studied in [34]–[36]. This form of the equations thus essentially corresponds to choosing bound (1.13).

1.8. The optimality conditions: reduced-order case. For design flexibility we also consider controllers of arbitrary reduced dimension. For the linear-quadratic problem without parameter uncertainty, the formulation of the necessary conditions given in [29] provides a generalization of LQG theory. Specifically, the optimal gains are characterized by a system of algebraic matrix equations consisting of a pair of modified Riccati equations and a pair of modified Lyapunov equations coupled by an oblique projection. When the order of the controller is equal to the order of the plant, the projection becomes the identity and the standard LQG result is recovered.

The outcome of the development above is a set of algebraic matrix equations that correspond to the necessary conditions for the Auxiliary Minimization Problem and hence to sufficient conditions for robust stability and performance. These necessary conditions characterize full- or reduced-order controllers with either choice of bounds (1.13) and (1.14) for each uncertain parameter. For control-system design, these equations can be used as follows. If a solution to the necessary conditions is obtained computationally and if certain definiteness conditions hold, *then* the explicitly synthesized controller (1) solves the Robust Stability Problem and (2) is guaranteed to provide robust performance bounded by $\operatorname{tr} \mathcal{P}\tilde{V}$ over the stipulated uncertainty range.

The applicability of these results is, of course, limited to plants that are nominally stabilizable via controllers of the given order. Indeed, in this case it has been shown [37] via topological degree theory that the optimality conditions for the case $\delta_i = 0$, $i = 1, \cdots, p$, possess at least one stabilizing solution. For the parameter uncertainty problem, i.e., $\delta_i > 0$, it follows from continuity properties that a solution also exists for sufficiently small δ_i. The *actual* range of uncertainty that can be stabilized and the tightness of the performance bound depend on the *conservatism* of our bounds. As will be seen from a numerical example, our bounds are not generally sharp. This is not unexpected, however, due to both the sense of the partial ordering employed in (1.12) and the fact that our choice of gain-invariant bounds permits a one-step, *noniterative* synthesis (rather than analysis) procedure. It should be noted that necessary and sufficient conditions for robust analysis of a block-structured class of uncertainties are obtainable using the μ-function [6]. This block structure, however, does not appear to include either the linear uncertainty model (1.1) or the matched uncertainty model of [21] as special cases.

In the present paper we present results of an illustrative numerical study for a well-known example used in [2] to demonstrate the lack of gain margin for LQG controllers. This type of uncertainty is a special case of (1.1) obtained by taking $p = m$ and defining B_i to be the matrix whose ith column is the same as the ith column of B, and zero otherwise. To obtain full-order, robustified controllers exhibiting performance/robustness tradeoffs, we use bound (1.13) for several values of δ_i. To obtain

these numerical results we used a straightforward iterative algorithm that requires only an LQG-type software package. The homotopy algorithm of [37] with appropriate extensions can also be used. Further descriptions of related algorithms and numerical results can be found in [38]-[40].

The development herein is self-contained, with the exception that the detailed derivation of the optimality conditions has been omitted. In specialized cases the derivation has been given previously. For the case of bound (1.13) only, a derivation using Kronecker products appears in [36]. Also, a derivation without parameter uncertainties has been given in [29] using Lagrange multipliers. Overall, the derivation involves considerable matrix manipulation. Since the detailed derivation does not appear to warrant the required space, we give an outline of the proof to assist the sufficiently motivated reader in reconstructing the details.

2. Notation and definitions. (Note that all matrices have real entries.)

$\mathbb{R}$, $\mathbb{R}^{r\times s}$, $\mathbb{R}^r$, $\mathbb{E}$	real numbers, $r\times s$ real matrices, $\mathbb{R}^{r\times 1}$, expectation
$\|\cdot\|$	Euclidean vector norm
I_r, $0_{r\times s}$, 0_r	$r\times r$ identity matrix, $r\times s$ zero matrix, $0_{r\times r}$
$(\,)^T$, $(\,)^{-1}$, $(\,)^{-T}$	transpose, inverse, inverse transpose
tr	trace
$\oplus$, $\otimes$	Kronecker sum, Kronecker product [41]
$\mathbb{S}^r$	$r\times r$ symmetric matrices
$\mathbb{N}^r$	$r\times r$ symmetric nonnegative-definite matrices
$\mathbb{P}^r$	$r\times r$ symmetric positive-definite matrices
$Z_1 \geqq Z_2$	$Z_1 - Z_2 \in \mathbb{N}^r$, $Z_1, Z_2 \in \mathbb{S}^r$
$Z_1 > Z_2$	$Z_1 - Z_2 \in \mathbb{P}^r$, $Z_1, Z_2 \in \mathbb{S}^r$
asymptotically stable matrix	matrix with eigenvalues in open left half-plane
n, m, l, p, n_c, n_i, m_i	positive integers, $i\in\{1,\cdots,p\}$
$\tilde{n}$, $\tilde{n}_i$	$n+n_c$, n_i+m_i, $i\in\{1,\cdots,p\}$
x, u, y, x_c	n, m, l, n_c-dimensional vectors
A, A_i; B, B_i; C, C_i	$n\times n$ matrices, $n\times m$ matrices, $l\times n$ matrices, $i\in\{1,\cdots,p\}$
A_c, B_c, C_c	$n_c\times n_c$, $n_c\times l$, $m\times n_c$ matrices
$\tilde{A}$, $\tilde{A}_i$	$\begin{bmatrix} A & BC_c \\ B_cC & A_c \end{bmatrix}$, $\begin{bmatrix} A_i & B_iC_c \\ B_cC_i & 0 \end{bmatrix}$, $i\in\{1,\cdots,p\}$
δ_i	positive number, $i\in\{1,\cdots,p\}$
Δ	$[-\delta_1,\delta_1]\times\cdots\times[-\delta_p,\delta_p]$
σ_i	real number, $i\in\{1,\cdots,p\}$
σ	$(\sigma_1,\cdots,\sigma_p)$
$\hat{\sigma}_i(\cdot)$	Lebesgue measurable function on $[0,\infty)$, $i\in\{1,\cdots,p\}$
$\hat{\sigma}(\cdot)$	$(\hat{\sigma}_1(\cdot),\cdots,\hat{\sigma}_p(\cdot))$
$L_\infty([0,\infty),\Delta)$	Lebesgue measurable functions on $[0,\infty)$ with values in Δ
α_i	positive number, $i\in\{1,\cdots,p\}$
D_i, E_i, H_i, K_i	$n\times n_i$, $n_i\times n$, $n\times m_i$, $m_i\times m$ matrices, $i\in\{1,\cdots,p\}$
$\tilde{D}_i$, $\tilde{E}_i$	$\tilde{n}\times\tilde{n}_i$, $\tilde{n}_i\times\tilde{n}$ matrices, $i\in\{1,\cdots,p\}$
$\sum'$, $\sum''$	see § 6
R_1	state weighting matrix in $\mathbb{N}^n$

R_2	control weighting matrix in $\mathbb{P}^m$
R_{12}	$n \times m$ cross weighting matrix such that $R_1 - R_{12}R_2^{-1}R_{12}^T \geqq 0$
$\tilde{R}$	$\begin{bmatrix} R_1 & R_{12}C_c \\ C_c^T R_{12}^T & C_c^T R_2 C_c \end{bmatrix}$
$w_1(\cdot)$	n-dimensional white noise
$w_2(\cdot)$	l-dimensional white noise
V_1	intensity of $w_1(\cdot)$ in $\mathbb{N}^n$
V_2	intensity of $w_2(\cdot)$ in $\mathbb{P}^l$
V_{12}	$n \times l$ cross-intensity of $w_1(\cdot)$, $w_2(\cdot)$
$\tilde{V}$	$\begin{bmatrix} V_1 & V_{12}B_c^T \\ B_c V_{12}^T & B_c V_2 B_c^T \end{bmatrix}$

3. Robust Stability and Robust Performance Problems. In this section we state the Robust Stability Problem and Robust Performance Problem along with related notation for later use.

3.1. Robust Stability Problem. For fixed $n_c \leqq n$, determine $(A_c, B_c, C_c) \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$ such that the closed-loop system consisting of the nth-order controlled plant

$$\dot{x}(t) = \left(A + \sum_{i=1}^{p} \hat{\sigma}_i(t) A_i\right) x(t) + \left(B + \sum_{i=1}^{p} \hat{\sigma}_i(t) B_i\right) u(t) \quad \text{a.a. } t \in [0, \infty), \tag{3.1}$$

measurements

$$y(t) = \left(C + \sum_{i=1}^{p} \hat{\sigma}_i(t) C_i\right) x(t), \tag{3.2}$$

and n_cth-order dynamic compensator

$$\dot{x}_c(t) = A_c x_c(t) + B_c y(t), \tag{3.3}$$

$$u(t) = C_c x_c(t) \tag{3.4}$$

are asymptotically stable[2] for all $\hat{\sigma}(\cdot) \in L_\infty([0, \infty), \Delta)$.

3.2. Robust Performance Problem. For fixed $n_c \leqq n$, determine $(A_c, B_c, C_c) \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$ such that, for the closed-loop system consisting of the nth-order controlled and disturbed plant

$$\dot{x}(t) = \left(A + \sum_{i=1}^{p} \hat{\sigma}_i(t) A_i\right) x(t) + \left(B + \sum_{i=1}^{p} \hat{\sigma}_i(t) B_i\right) u(t) + w_1(t) \quad \text{a.a. } t \in [0, \infty), \tag{3.5}$$

noisy measurements

$$y(t) = \left(C + \sum_{i=1}^{p} \hat{\sigma}_i(t) C_i\right) x(t) + w_2(t), \tag{3.6}$$

and n_cth-order dynamic compensator (3.3), (3.4), the performance criterion

$$J(A_c, B_c, C_c) \triangleq \sup_{\hat{\sigma}(\cdot) \in L_\infty([0,\infty),\Delta)} \limsup_{t \to \infty} \mathbb{E}[x^T(t) R_1 x(t) + 2x^T(t) R_{12} u(t) + u^T(t) R_2 u(t)] \tag{3.7}$$

is minimized.

[2] Asymptotic stability for a nonautonomous system is defined in the standard way (see, e.g., [42]).

For each controller (A_c, B_c, C_c) and parameter variation $\hat{\sigma}(\cdot) \in L_\infty([0, \infty), \Delta)$ the undisturbed closed-loop system (3.1)-(3.4) is given by

$$\dot{\tilde{x}}(t) = \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right) \tilde{x}(t) \quad \text{a.a. } t \in [0, \infty), \tag{3.8}$$

while the disturbed closed-loop system (3.3)-(3.6) is

$$\dot{\tilde{x}}(t) = \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right) \tilde{x}(t) + \tilde{w}(t) \quad \text{a.a. } t \in [0, \infty). \tag{3.9}$$

Also (see, e.g., [43, p. 194]), let $\tilde{\Phi} : [0, \infty) \to \mathbb{R}^{\tilde{n} \times \tilde{n}}$ be the unique absolutely continuous solution to

$$\dot{\tilde{\Phi}}(t) = \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right) \tilde{\Phi}(t) \quad \text{a.a. } t \in [0, \infty), \tag{3.10}$$

$$\tilde{\Phi}(0) = I_{\tilde{n}}, \tag{3.11}$$

and recall that $\tilde{\Phi}^{-1}(\cdot)$ satisfies

$$\frac{d}{dt} \tilde{\Phi}^{-1}(t) = -\tilde{\Phi}^{-1}(t) \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right) \quad \text{a.a. } t \in [0, \infty). \tag{3.12}$$

4. Sufficient conditions for robust stability and performance. For robust stability we characterize quadratic Lyapunov functions for the closed-loop system.

THEOREM 4.1. *Let* $\Omega : \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c} \to \mathbb{S}^{\tilde{n}}$ *satisfy*

$$\begin{aligned} &\sum_{i=1}^{p} \sigma_i (\tilde{A}_i^T \mathcal{P} + \mathcal{P} \tilde{A}_i) < \Omega(\mathcal{P}, B_c, C_c), \qquad \sigma \in \Delta, \\ &(\mathcal{P}, B_c, C_c) \in \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}. \end{aligned} \tag{4.1}$$

If, for some $(A_c, B_c, C_c) \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$, *there exists* $\mathcal{P} \in \mathbb{P}^{\tilde{n}}$ *satisfying*

$$0 = \tilde{A}^T \mathcal{P} + \mathcal{P} \tilde{A} + \Omega(\mathcal{P}, B_c, C_c), \tag{4.2}$$

then (A_c, B_c, C_c) *solves the Robust Stability Problem.*

Proof. Define the Lyapunov function

$$V(\tilde{x}) \triangleq \tilde{x}^T \mathcal{P} \tilde{x}, \qquad \tilde{x} \in \mathbb{R}^{\tilde{n}}.$$

For almost all $t \in [0, \infty)$ and $\tilde{x}(t)$ satisfying (3.8), it follows from (4.2) that

$$\begin{aligned} \dot{V}(\tilde{x}(t)) &= \dot{\tilde{x}}^T(t) \mathcal{P} \tilde{x}(t) + \tilde{x}^T(t) \mathcal{P} \dot{\tilde{x}}(t) \\ &= \tilde{x}^T(t) \left[\left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right)^T \mathcal{P} + \mathcal{P} \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t) \tilde{A}_i \right) \right] \tilde{x}(t) \\ &= \tilde{x}^T(t) \left[\sum_{i=1}^{p} \hat{\sigma}_i(t) (\tilde{A}_i^T \mathcal{P} + \mathcal{P} \tilde{A}_i) - \Omega(\mathcal{P}, B_c, C_c) \right] \tilde{x}(t). \end{aligned}$$

Since $\hat{\sigma}(t) \in \Delta$, almost all $t \in [0, \infty)$, it follows from (4.1) that there exists $\gamma > 0$ such that $\dot{V}(\tilde{x}(t)) \leqq -\gamma \|\tilde{x}(t)\|^2$, almost all $t \in [0, \infty)$. □

Remark 4.1. If (A_c, B_c, C_c) solves the Robust Stability Problem, then

$$\lim_{t \to \infty} \tilde{\Phi}(t) = 0, \qquad \hat{\sigma}(\cdot) \in L_\infty([0, \infty), \Delta). \tag{4.3}$$

Remark 4.2. As will be seen, the bound (4.1) will be guaranteed for all $\mathcal{P}$, B_c, C_c by suitable construction of the function Ω. In addition, the existence of a solution $\mathcal{P}$ to (4.2) need not be verified in practice. Rather, (4.2) is a result of numerically solving the necessary conditions for the Auxiliary Minimization Problem given in Theorem 6.1.

For the Robust Performance Problem the cost can be expressed in terms of the closed-loop second-moment matrix.

PROPOSITION 4.1. *For* $(A_c, B_c, C_c) \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$ *and* $\hat{\sigma}(\cdot) \in L_\infty([0, \infty), \Delta)$ *the second-moment matrix*

$$\tilde{Q}(t) \triangleq \mathbb{E}[\tilde{x}(t)\tilde{x}^T(t)], \qquad t \in [0, \infty), \tag{4.4}$$

satisfies

$$\dot{\tilde{Q}}(t) = \left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t)\tilde{A}_i\right)\tilde{Q}(t) + \tilde{Q}(t)\left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t)\tilde{A}_i\right)^T + \tilde{V} \quad \text{a.a. } t \in [0, \infty), \tag{4.5}$$

or, equivalently,

$$\tilde{Q}(t) = \tilde{\Phi}(t)\tilde{Q}(0)\tilde{\Phi}^T(t) + \int_0^t \tilde{\Phi}(t)\tilde{\Phi}^{-1}(s)\tilde{V}\tilde{\Phi}^{-T}(s)\tilde{\Phi}^T(t)\, ds, \qquad t \in [0, \infty). \tag{4.6}$$

Furthermore,

$$J(A_c, B_c, C_c) = \sup_{\hat{\sigma}(\cdot) \in L_\infty([0,\infty),\Delta)} \limsup_{t \to \infty} \operatorname{tr} \tilde{Q}(t)\tilde{R}, \tag{4.7}$$

or, equivalently,

$$\begin{aligned} J(A_c, B_c, C_c) \triangleq \sup_{\hat{\sigma}(\cdot) \in L_\infty([0,\infty),\Delta)} \limsup_{t \to \infty} \operatorname{tr} \Bigg[& \tilde{\Phi}(t)\tilde{Q}(0)\tilde{\Phi}^T(t)\tilde{R} \\ & + \int_0^t \tilde{\Phi}(t)\tilde{\Phi}^{-1}(s)\tilde{V}\tilde{\Phi}^{-T}(s)\tilde{\Phi}^T(t)\, ds\, \tilde{R} \Bigg]. \end{aligned} \tag{4.8}$$

Proof. The second-moment equation (4.5) is a direct consequence of the Itô differential rule (see [44, p. 142]), while (4.6) follows by direct verification. Finally, (4.7) is immediate. □

We now obtain an upper bound for J in terms of the matrix $\mathscr{P}$. The following lemma is required.

LEMMA 4.1. *Let* $\Omega : \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c} \to \mathbb{S}^{\tilde{n}}$ *and* $(A_c, B_c, C_c) \in \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$ *be given. Then* $\mathscr{P} \in \mathbb{P}^{\tilde{n}}$ *satisfies* (4.2) *if and only if* $\mathscr{P}$ *satisfies*

$$\begin{aligned} \mathscr{P} = \tilde{\Phi}^T(t)\mathscr{P}\tilde{\Phi}(t) + \int_0^t \tilde{\Phi}^T(t)\tilde{\Phi}^{-T}(s) \\ \cdot \left[\Omega(\mathscr{P}, B_c, C_c) - \sum_{i=1}^{p} \hat{\sigma}_i(t)(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i)\right]\tilde{\Phi}^{-1}(s)\tilde{\Phi}(t)\, ds, \\ \hat{\sigma}(\cdot) \in L_\infty([0, \infty), \Delta), \qquad t \in [0, \infty). \end{aligned} \tag{4.9}$$

Proof. Suppose $\mathscr{P}$ satisfies (4.2). Then for $t \in [0, \infty)$,

$$\begin{aligned} 0 &= \tilde{\Phi}^{-T}(t)\left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t)\tilde{A}_i\right)^T \mathscr{P}\tilde{\Phi}^{-1}(t) + \tilde{\Phi}^{-T}(t)\mathscr{P}\left(\tilde{A} + \sum_{i=1}^{p} \hat{\sigma}_i(t)\tilde{A}_i\right)\tilde{\Phi}^{-1}(t) \\ &\quad + \tilde{\Phi}^{-T}(t)\left[\Omega(\mathscr{P}, B_c, C_c) - \sum_{i=1}^{p} \hat{\sigma}_i(t)(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i)\right]\tilde{\Phi}^{-1}(t) \\ &= -\frac{d}{dt}\left[\tilde{\Phi}^{-T}(t)\mathscr{P}\tilde{\Phi}^{-1}(t)\right] + \tilde{\Phi}^{-T}(t)\left[\Omega(\mathscr{P}, B_c, C_c) - \sum_{i=1}^{p} \hat{\sigma}_i(t)(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i)\right]\tilde{\Phi}^{-1}(t), \end{aligned}$$

which yields

$$\begin{aligned} 0 = -\tilde{\Phi}^{-T}(t)\mathscr{P}\tilde{\Phi}^{-1}(t) + \mathscr{P} \\ + \int_0^t \tilde{\Phi}^{-T}(s)\left[\Omega(\mathscr{P}, B_c, C_c) - \sum_{i=1}^{p} \hat{\sigma}_i(s)(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i)\right]\tilde{\Phi}^{-1}(s)\, ds. \end{aligned}$$

Thus (4.9) is satisfied. Conversely, suppose $\mathscr{P}$ satisfies (4.9). Differentiating with respect to t using Leibniz's rule yields

$$
\begin{aligned}
0 &= \Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)^T\tilde{\Phi}^T(t)\mathscr{P}\tilde{\Phi}(t)+\tilde{\Phi}^T(t)\mathscr{P}\tilde{\Phi}(t)\Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)\\
&\quad+\Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)^T\int_0^t\tilde{\Phi}^T(t)\tilde{\Phi}^{-T}(s)\\
&\quad\cdot\Big[\Omega(\mathscr{P},B_c,C_c)-\sum_{i=1}^{p}\hat{\sigma}_i(s)(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)\Big]\tilde{\Phi}^{-1}(s)\tilde{\Phi}(t)\,ds\\
&\quad+\int_0^t\tilde{\Phi}^T(t)\tilde{\Phi}^{-T}(s)\Big[\Omega(\mathscr{P},B_c,C_c)-\sum_{i=1}^{p}\hat{\sigma}_i(s)(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)\Big]\\
&\quad\cdot\tilde{\Phi}^{-1}(s)\tilde{\Phi}(t)\,ds\Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)\\
&\quad+\Omega(\mathscr{P},B_c,C_c)-\sum_{i=1}^{p}\hat{\sigma}_i(t)(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)\\
&=\Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)^T\mathscr{P}+\mathscr{P}\Big(\tilde{A}+\sum_{i=1}^{p}\hat{\sigma}_i(t)\tilde{A}_i\Big)+\Omega(\mathscr{P},B_c,C_c)-\sum_{i=1}^{p}\hat{\sigma}_i(t)(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)\\
&=\tilde{A}^T\mathscr{P}+\mathscr{P}\tilde{A}+\Omega(\mathscr{P},B_c,C_c).
\end{aligned}
$$

Hence (4.2) is satisfied. □

Remark 4.3. Note the identity

$$
\begin{aligned}
&\operatorname{tr}\int_0^t\tilde{\Phi}(t)\tilde{\Phi}^{-1}(s)\tilde{V}\tilde{\Phi}^{-T}(s)\tilde{\Phi}^T(t)\,ds\,\tilde{R}=\operatorname{tr}\int_0^t\tilde{\Phi}^T(t)\tilde{\Phi}^{-T}(s)\tilde{R}\tilde{\Phi}^{-1}(s)\tilde{\Phi}(t)\,ds\,\tilde{V},\\
&(A_c,B_c,C_c)\in\mathbb{R}^{n_c\times n_c}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c},\quad\hat{\sigma}(\cdot)\in L_\infty([0,\infty),\Delta),\quad t\in[0,\infty).
\end{aligned}
\tag{4.10}
$$

We are now in a position to bound the cost J by means of the matrix $\mathscr{P}$.

THEOREM 4.2. *Let* $\Omega:\mathbb{P}^{\tilde{n}}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c}\to\mathbb{S}^{\tilde{n}}$ *satisfy* (4.1) *and*

$$
\begin{aligned}
&\sum_{i=1}^{p}\sigma_i(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)+\tilde{R}\leqq\Omega(\mathscr{P},B_c,C_c),\qquad\sigma\in\Delta,\\
&(\mathscr{P},B_c,C_c)\in\mathbb{P}^{\tilde{n}}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c}.
\end{aligned}
\tag{4.11}
$$

If, for some $(A_c,B_c,C_c)\in\mathbb{R}^{n_c\times n_c}\times\mathbb{R}^{n_c\times l}\times\mathbb{R}^{m\times n_c}$, *there exists* $\mathscr{P}\in\mathbb{P}^{\tilde{n}}$ *satisfying* (4.2), *then*

$$
J(A_c,B_c,C_c)\leqq\operatorname{tr}\mathscr{P}\tilde{V}.
\tag{4.12}
$$

Proof. From (4.8)–(4.10) and (4.3) it follows that

$$
\begin{aligned}
&J(A_c,B_c,C_c)\\
&=\sup_{\hat{\sigma}(\cdot)\in L_\infty([0,\infty),\Delta)}\limsup_{t\to\infty}\operatorname{tr}\Big\{\tilde{\Phi}(t)\tilde{Q}(0)\tilde{\Phi}^T(t)\tilde{R}+\mathscr{P}\tilde{V}-\tilde{\Phi}^T(t)\mathscr{P}\tilde{\Phi}(t)\tilde{V}\\
&\quad-\int_0^t\tilde{\Phi}^T(t)\tilde{\Phi}^{-T}(s)\Big[\Omega(\mathscr{P},B_c,C_c)-\tilde{R}-\sum_{i=1}^{p}\hat{\sigma}_i(s)(\tilde{A}_i^T\mathscr{P}+\mathscr{P}\tilde{A}_i)\Big]\tilde{\Phi}^{-1}(s)\tilde{\Phi}(t)\,ds\,\tilde{V}\Big\}\\
&\leqq\sup_{\hat{\sigma}(\cdot)\in L_\infty([0,\infty),\Delta)}\limsup_{t\to\infty}\operatorname{tr}[\tilde{\Phi}(t)\tilde{Q}(0)\tilde{\Phi}^T(t)\tilde{R}+\mathscr{P}\tilde{V}]\\
&=\operatorname{tr}\mathscr{P}\tilde{V}.
\end{aligned}
$$

□

Remark 4.4. Note that since $\tilde{R} \geqq 0$, (4.11) implies

$$\sum_{i=1}^{p} \sigma_i(\tilde{A}_i^T \mathscr{P} + \mathscr{P}\tilde{A}_i) \leqq \Omega(\mathscr{P}, B_c, C_c), \qquad \sigma \in \Delta, \tag{4.13}$$

which is a weak form of (4.1). If $\tilde{R} > 0$ then (4.11) implies (4.1). This implication is not surprising since (4.11) implies robust performance while (4.1) implies robust stability.

5. Choice of bounds. To satisfy (4.11), $\Omega(\cdot, \cdot, \cdot)$ is chosen to be of the form

$$\Omega(\mathscr{P}, B_c, C_c) = \sum_{i=1}^{p} \Lambda_i(\mathscr{P}, B_c, C_c) + \tilde{R}, \tag{5.1}$$

where, for each $i = 1, \cdots, p$, $\Lambda_i : \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c} \to \mathbb{S}^{\tilde{n}}$ satisfies

$$\begin{aligned} &\sigma_i(\tilde{A}_i^T \mathscr{P} + \mathscr{P}\tilde{A}_i) \leqq \Lambda_i(\mathscr{P}, B_c, C_c), \qquad \sigma_i \in [-\delta_i, \delta_i], \\ &(\mathscr{P}, B_c, C_c) \in \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}. \end{aligned} \tag{5.2}$$

Two distinct choices for the bound Λ_i are considered. As we pointed out in § 1, the first choice corresponds to a right shift/multiplicative white noise model [33], while the second bound generalizes results found in [26].

PROPOSITION 5.1. *For all $\alpha_i > 0$ the function*

$$\Lambda_i(\mathscr{P}, B_c, C_c) = \delta_i(\alpha_i \mathscr{P} + \alpha_i^{-1} \tilde{A}_i^T \mathscr{P} \tilde{A}_i) \tag{5.3}$$

satisfies (5.2).

Proof. Note that

$$\begin{aligned} 0 &\leqq [\sigma_i(\alpha_i/\delta_i)^{1/2} I_{\tilde{n}} - (\delta_i/\alpha_i)^{1/2} \tilde{A}_i]^T \mathscr{P} [\sigma_i(\alpha_i/\delta_i)^{1/2} I_{\tilde{n}} - (\delta_i/\alpha_i)^{1/2} \tilde{A}_i] \\ &= \sigma_i^2(\alpha_i/\delta_i)\mathscr{P} + (\delta_i/\alpha_i)\tilde{A}_i^T \mathscr{P} \tilde{A}_i - \sigma_i(\tilde{A}_i^T \mathscr{P} + \mathscr{P}\tilde{A}_i), \end{aligned}$$

which, since $\sigma_i^2 \leqq \delta_i^2$, implies (5.2). □

PROPOSITION 5.2. *For all $\tilde{D}_i \in \mathbb{R}^{\tilde{n} \times \tilde{n}_i}$ and $\tilde{E}_i \in \mathbb{R}^{\tilde{n}_i \times \tilde{n}}$ satisfying*

$$\tilde{A}_i = \tilde{D}_i \tilde{E}_i, \tag{5.4}$$

the function

$$\Lambda_i(\mathscr{P}, B_c, C_c) = \delta_i(\tilde{E}_i^T \tilde{E}_i + \mathscr{P}\tilde{D}_i \tilde{D}_i^T \mathscr{P}) \tag{5.5}$$

satisfies (5.2).

Proof. Note that

$$\begin{aligned} 0 &\leqq [\delta_i^{1/2} \tilde{E}_i - \sigma_i \delta_i^{-1/2} \tilde{D}_i^T \mathscr{P}]^T [\delta_i^{1/2} \tilde{E}_i - \sigma_i \delta_i^{-1/2} \tilde{D}_i^T \mathscr{P}] \\ &= \delta_i \tilde{E}_i^T \tilde{E}_i + (\sigma_i^2/\delta_i)\mathscr{P}\tilde{D}_i \tilde{D}_i^T \mathscr{P} - \sigma_i(\tilde{A}_i^T \mathscr{P} + \mathscr{P}\tilde{A}_i), \end{aligned}$$

which implies (5.2). □

6. The auxiliary minimization problem and necessary conditions for optimality. To optimize robust performance while retaining robust stability, we consider the following problem for which the cost functional is given by the bound (4.12).

6.1. Auxiliary Minimization Problem. For $i = 1, \cdots, p$, let Λ_i be given by either (5.3) or (5.5). Determine $(\mathscr{P}, A_c, B_c, C_c) \in \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times n_c} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$, which minimizes

$$\mathscr{J}(\mathscr{P}, A_c, B_c, C_c) \triangleq \operatorname{tr} \mathscr{P}\tilde{V} \tag{6.1}$$

subject to

$$0 = \tilde{A}^T\mathscr{P} + \mathscr{P}\tilde{A} + \sum_{i=1}^{p} \Lambda_i(\mathscr{P}, B_c, C_c) + \tilde{R}, \tag{6.2}$$

$$\sum_{i=1}^{p} \sigma_i(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i) < \sum_{i=1}^{p} \Lambda_i(\mathscr{P}, B_c, C_c) + \tilde{R}, \qquad \sigma \in \Delta. \tag{6.3}$$

Remark 6.1. Note that (6.3) enforces both (4.1) and (4.11) to guarantee robust stability and performance.

To derive first-order necessary conditions for the Auxiliary Minimization Problem, note that the constraint (6.3) defines an open set.

PROPOSITION 6.1. *The set of* $(\mathscr{P}, B_c, C_c) \in \mathbb{P}^{\tilde{n}} \times \mathbb{R}^{n_c \times l} \times \mathbb{R}^{m \times n_c}$ *satisfying* (6.3) *is open.*

Proof. Since $\Lambda_i(\cdot, \cdot, \cdot)$ is continuous it can be shown that the function

$$f(\mathscr{P}, B_c, C_c) \triangleq \min_{\sigma \in \Delta} \lambda_{\min} \left\{ \sum_{i=1}^{p} \Lambda_i(\mathscr{P}, B_c, C_c) + \tilde{R} - \sum_{i=1}^{p} \sigma_i(\tilde{A}_i^T\mathscr{P} + \mathscr{P}\tilde{A}_i) \right\}$$

is also continuous. Since (6.3) is equivalent to $0 < f(\mathscr{P}, B_c, C_c)$, the result is immediate. □

To obtain explicit feedback gain expressions we shall require two additional technical assumptions. If bound (5.3) is chosen for a given $i \in \{1, \cdots, p\}$ we require

$$B_i \neq 0 \Rightarrow C_i = 0, \tag{6.4}$$

i.e., B_i and C_i are not simultaneously nonzero. Of course, both B_i and C_i may be zero. Assumption (6.4) implies that parameter uncertainties in B and C must be modeled as uncorrelated. Correlation between uncertainties in A and B or A and C is, of course, permitted. Furthermore, if bound (5.5) is chosen for a given $i \in \{1, \cdots, p\}$ we require

$$C_i = 0. \tag{6.5}$$

We stress that (6.4) and (6.6) can be removed, but at the expense of explicit gain expressions.

When we use bound (5.3) the positive constant α_i will be considered fixed but arbitrary. Furthermore, for bound (5.5), let $D_i \in \mathbb{R}^{n \times n_i}$, $E_i \in \mathbb{R}^{n_i \times n}$, $H_i \in \mathbb{R}^{n \times m_i}$, and $K_i \in \mathbb{R}^{m_i \times m}$ satisfy

$$A_i = D_iE_i, \qquad B_i = H_iK_i, \tag{6.6}$$

and define $\tilde{D}_i$, $\tilde{E}_i$ satisfying (5.4) by

$$\tilde{D}_i \triangleq \begin{bmatrix} D_i & H_i \\ 0_{n_c \times n_i} & 0_{n_c \times m_i} \end{bmatrix}, \qquad \tilde{E}_i \triangleq \begin{bmatrix} E_i & 0_{n_i \times n_c} \\ 0_{m_i \times n} & K_iC_c \end{bmatrix}. \tag{6.7}$$

In addition to the open set defined by (6.3), the derivation of the necessary conditions requires that $(\mathscr{P}, A_c, B_c, C_c)$ be further restricted so that

$$\left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i\alpha_i I_{\tilde{n}} + \sum{}'' \delta_i\tilde{D}_i\tilde{D}_i^T\mathscr{P}\right) \oplus \left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i\alpha_i I_{\tilde{n}} + \sum{}'' \delta_i\tilde{D}_i\tilde{D}_i^T\mathscr{P}\right) + \sum{}' (\delta_i\alpha_i^{-1})\tilde{A}_i \otimes \tilde{A}_i \quad \text{is asymptotically stable,} \tag{6.8}$$

$$(A_c, B_c, C_c) \text{ is controllable and observable.} \tag{6.9}$$

In (6.8) the notation $\sum'$ and $\sum''$ denotes summation over indices for which bounds (5.3) and (5.5), respectively, have been chosen. Note that (6.8) and (6.9) play no role in the Auxiliary Minimization Problem and thus need not be verified for robust stability or robust performance.

For arbitrary $Q, P, \hat{Q}, \hat{P} \in \mathbb{R}^{n\times n}$ define the following notation:

$$R_{2a} \triangleq R_2 + \sum{}' (\delta_i \alpha_i^{-1}) B_i^T (P+\hat{P}) B_i + \sum{}'' \delta_i K_i^T K_i,$$

$$V_{2a} \triangleq V_2 + \sum{}' (\delta_i \alpha_i^{-1}) C_i (Q+\hat{Q}) C_i^T,$$

$$P_a \triangleq B^T P + R_{12}^T + \sum{}' (\delta_i \alpha_i^{-1}) B_i^T (P+\hat{P}) A_i,$$

$$Q_a \triangleq QC^T + V_{12} + \sum{}' (\delta_i \alpha_i^{-1}) A_i (Q+\hat{Q}) C_i^T,$$

$$D \triangleq \sum{}'' \delta_i (D_i D_i^T + H_i H_i^T), \qquad E \triangleq \sum{}'' \delta_i E_i^T E_i,$$

$$\hat{A} \triangleq A + \frac{1}{2} \sum{}' \delta_i \alpha_i I_n, \quad \hat{A}_P \triangleq \hat{A} - B R_{2a}^{-1} P_a, \quad \hat{A}_Q \triangleq \hat{A} - Q_a V_{2a}^{-1} C.$$

The following lemma will be needed.

LEMMA 6.1. *If* $\hat{Q}, \hat{P} \in \mathbb{N}^n$ *and* rank $\hat{Q}\hat{P} = n_c$, *then there exist* $G, \Gamma \in \mathbb{R}^{n_c \times n}$ *and invertible* $M \in \mathbb{R}^{n_c \times n_c}$ *such that*

$$\hat{Q}\hat{P} = G^T M \Gamma, \tag{6.10}$$

$$\Gamma G^T = I_{n_c}. \tag{6.11}$$

Furthermore, G, M, and Γ *are unique except for a change of basis in* $\mathbb{R}^{n_c}$.

Proof. The result is an immediate consequence of [45, Thm. 6.2.5, p. 123]. □

Note that because of (6.11), the $n \times n$ matrix $\tau \triangleq G^T \Gamma$ is idempotent, i.e., $\tau^2 = \tau$. Since τ is not necessarily symmetric, it is an oblique projection. Also, define $\tau_\perp \triangleq I_n - \tau$.

THEOREM 6.1. *Suppose* $(\mathscr{P}, A_c, B_c, C_c)$ *solves the Auxiliary Minimization Problem subject to* (6.8) *and* (6.9). *Then there exist* $P, Q, \hat{P}, \hat{Q} \in \mathbb{N}^n$ *such that* $\mathscr{P}, A_c, B_c, C_c$ *are given by*

$$\mathscr{P} = \begin{bmatrix} P+\hat{P} & -\hat{P}G^T \\ -G\hat{P} & G\hat{P}G^T \end{bmatrix}, \tag{6.12}$$

$$A_c = \Gamma (A - Q_a V_{2a}^{-1} C - B R_{2a}^{-1} P_a + DP) G^T, \tag{6.13}$$

$$B_c = \Gamma Q_a V_{2a}^{-1}, \tag{6.14}$$

$$C_c = -R_{2a}^{-1} P_a G^T, \tag{6.15}$$

and such that $P, Q, \hat{P}, \hat{Q}$ *satisfy*

$$\begin{aligned} 0 = {} & \hat{A}^T P + P\hat{A} + R_1 + \sum{}' (\delta_i \alpha_i^{-1}) [A_i^T P A_i + (A_i - Q_a V_{2a}^{-1} C_i)^T \hat{P} (A_i - Q_a V_{2a}^{-1} C_i)] \\ & + E + PDP - P_a^T R_{2a}^{-1} P_a + \tau_\perp^T P_a^T R_{2a}^{-1} P_a \tau_\perp, \end{aligned} \tag{6.16}$$

$$\begin{aligned} 0 = {} & [\hat{A} + D(P+\hat{P})] Q + Q[\hat{A} + D(P+\hat{P})]^T + V_1 \\ & + \sum{}' (\delta_i \alpha_i^{-1}) [A_i Q A_i^T + (A_i - B_i R_{2a}^{-1} P_a) \hat{Q} (A_i - B_i R_{2a}^{-1} P_a)^T] \\ & - Q_a V_{2a}^{-1} Q_a^T + \tau_\perp Q_a V_{2a}^{-1} Q_a^T \tau_\perp^T, \end{aligned} \tag{6.17}$$

$$0 = (\hat{A}_Q + DP)^T \hat{P} + \hat{P}(\hat{A}_Q + DP) + \hat{P} D \hat{P} + P_a^T R_{2a}^{-1} P_a - \tau_\perp^T P_a^T R_{2a}^{-1} P_a \tau_\perp, \tag{6.18}$$

$$0 = (\hat{A}_P + DP)\hat{Q} + \hat{Q}(\hat{A}_P + DP)^T + Q_a V_{2a}^{-1} Q_a^T - \tau_\perp Q_a V_{2a}^{-1} Q_a^T \tau_\perp^T, \tag{6.19}$$

$$\text{rank } \hat{Q} = \text{rank } \hat{P} = \text{rank } \hat{Q}\hat{P} = n_c. \tag{6.20}$$

Conversely, if there exist $P, Q, \hat{P}, \hat{Q} \in \mathbb{N}^n$ *satisfying* (6.16)-(6.20), *then* $\mathscr{P}$ *given by* (6.12) *satisfies* (6.2) *or, equivalently,* (4.2) *with* (A_c, B_c, C_c) *given by* (6.13)-(6.15).

Outline of proof. As discussed in § 1, we limit the presentation of the proof to the salient details. First note that with the choice of bounds Λ_i, (6.2) becomes

$$\begin{aligned} 0 = &\left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i \alpha_i I_{\tilde{n}}\right)^T \mathscr{P} + \mathscr{P}\left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i \alpha_i I_{\tilde{n}}\right) + \tilde{R} \\ &+ \sum{}' (\delta_i \alpha_i^{-1}) \tilde{A}_i^T \mathscr{P} \tilde{A}_i + \sum{}'' \delta_i (\tilde{E}_i^T \tilde{E}_i + \mathscr{P} \tilde{D}_i \tilde{D}_i^T \mathscr{P}). \end{aligned} \tag{6.21}$$

By introducing multipliers $\lambda \in \mathbb{R}$, $\lambda \geqq 0$, and $\mathscr{Q} \in \mathbb{R}^{\tilde{n} \times \tilde{n}}$, a Lagrangian can be defined as

$$\mathscr{L}(\mathscr{P}, A_c, B_c, C_c) \triangleq \operatorname{tr}[\lambda \mathscr{P} \tilde{V} + \mathscr{Q}(\text{RHS of (6.21)})]. \tag{6.22}$$

Setting $\partial \mathscr{L}/\partial \mathscr{P} = 0$ and using (6.8) implies that $\lambda = 1$ without loss of generality, $\mathscr{Q} \geqq 0$, and $\mathscr{Q}$ satisfies

$$\begin{aligned} 0 = &\left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i \alpha_i I_{\tilde{n}} + \sum{}'' \tilde{D}_i \tilde{D}_i^T \mathscr{P}\right) \mathscr{Q} + \mathscr{Q}\left(\tilde{A} + \frac{1}{2}\sum{}' \delta_i \alpha_i I_{\tilde{n}} + \sum{}'' \tilde{D}_i \tilde{D}_i^T \mathscr{P}\right)^T \\ &+ \sum{}' (\delta_i \alpha_i^{-1}) \tilde{A}_i \mathscr{Q} \tilde{A}_i^T + \tilde{V}. \end{aligned} \tag{6.23}$$

The remainder of the derivation is exactly parallel to the techniques utilized in [29] and [36]. Briefly, the principal steps are as follows:

Step 1. Compute $\partial \mathscr{L}/\partial A_c$, $\partial \mathscr{L}/\partial B_c$, and $\partial \mathscr{L}/\partial C_c$.
Step 2. Use (6.9) to show that the lower right $n_c \times n_c$ blocks of $\mathscr{Q}$ and $\mathscr{P}$ are positive definite.
Step 3. Use $\partial \mathscr{L}/\partial A_c = 0$ to define a projection τ and new variables $P, Q, \hat{P}, \hat{Q}, G, \Gamma$.
Step 4. Partition (6.21) and (6.23) into six equations (1)-(6) corresponding to the $n \times n$, $n \times n_c$ and $n_c \times n_c$ blocks of $\mathscr{P}$ and $\mathscr{Q}$, respectively.
Step 5. Use (2) and (3) to solve for A_c; show that (5) and (6) also yield A_c; note that with A_c now given, (3) and (6) are superfluous and can be eliminated.
Step 6. Manipulate (1), (2), (4), and (5) to yield (6.16)-(6.19).
Step 7. Show that Steps 4-6 are reversible so that (6.16)-(6.20) are equivalent to (6.2) or, equivalently, (4.2).

By enforcing the strict inequalities $\mathscr{P} > 0$ and (6.3), solutions of (6.16)-(6.20) guarantee robust stability with a robust performance bound. The following result follows from Theorem 4.1, Theorem 4.2, and the converse of Theorem 6.1.

THEOREM 6.2. *Suppose there exist* $P, Q, \hat{P}, \hat{Q} \in \mathbb{N}^n$ *satisfying* (6.16)-(6.20), *and suppose that* (6.3) *and* $\mathscr{P} > 0$ *are satisfied with* $(\mathscr{P}, A_c, B_c, C_c)$ *given by* (6.12)-(6.15). *Then the compensator* A_c, B_c, C_c *given by* (6.13)-(6.15) *solves the Robust Stability Problem and the closed-loop performance* (3.7) *satisfies the bound*

$$J(A_c, B_c, C_c) \leqq \operatorname{tr} \mathscr{P} \tilde{V}. \tag{6.24}$$

The following existence result concerns the solvability of (6.16)-(6.20). Let n_u denote the dimension of the unstable subspace of the plant dynamics matrix A.

THEOREM 6.3. *Assuming* $n_c \geqq n_u$, $R_1 > 0$, $V_1 > 0$, *suppose the nominal plant, i.e.,* (3.1), (3.2) *with* $\delta_i = 0$, $i = 1, \cdots, p$, *is stabilizable and detectable and, in addition, is stabilizable by means of an* n_c*th-order strictly proper dynamic compensator* (3.3), (3.4). *Then there exist* $\bar{\delta}_1, \cdots, \bar{\delta}_p > 0$ *such that if* $\delta_i \in [0, \bar{\delta}_i)$, $i = 1, \cdots, p$, *then* (6.16)-(6.20) *have a solution* $P, Q, \hat{P}, \hat{Q} \in \mathbb{N}^n$ *for which* (A_c, B_c, C_c) *given by* (6.13)-(6.15) *solve the robust stability problem with robust performance bound* (6.24).

Proof. From Theorem 3.1 of [37] it follows that there exists a solution to (6.16)-(6.20) that stabilizes the nominal plant. By continuity there exists a neighborhood over which robust stability with performance bound (6.24) holds. □

Theorem 6.3 is an existence result that guarantees solvability of the sufficiency conditions over a range of parameter uncertainties. The actual range of uncertainty that can be bounded and the conservatism of the performance bound are problem dependent. To this end we now consider a numerical example.

7. Illustrative numerical example. To demonstrate the theory above we present an illustrative numerical example. The example chosen was originally used in [2] to illustrate the lack of a guaranteed gain margin for LQG controllers. This example was also considered in [35] for a preliminary robustness study and reconsidered in [46] using μ-analysis. Define the following:

$$n = n_u = 2, \qquad m = l = p = 1,$$

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = [1 \quad 0],$$

$$A_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C_1 = [0 \quad 0],$$

$$R_1 = V_1 = \begin{bmatrix} 60 & 60 \\ 60 & 60 \end{bmatrix}, \quad R_{12} = V_{12} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad R_2 = V_2 = 1.$$

Note that the system is open-loop unstable and becomes uncontrollable at $\sigma_1 = -1$. As can be seen using root locus, a strictly proper stabilizing controller must be of at least second order. Hence we consider (6.16)-(6.20) with $n_c = n$ and $\tau_\perp = 0$. Furthermore, we use bound (5.3) and thus set $D = E = 0$. Using algorithms described in

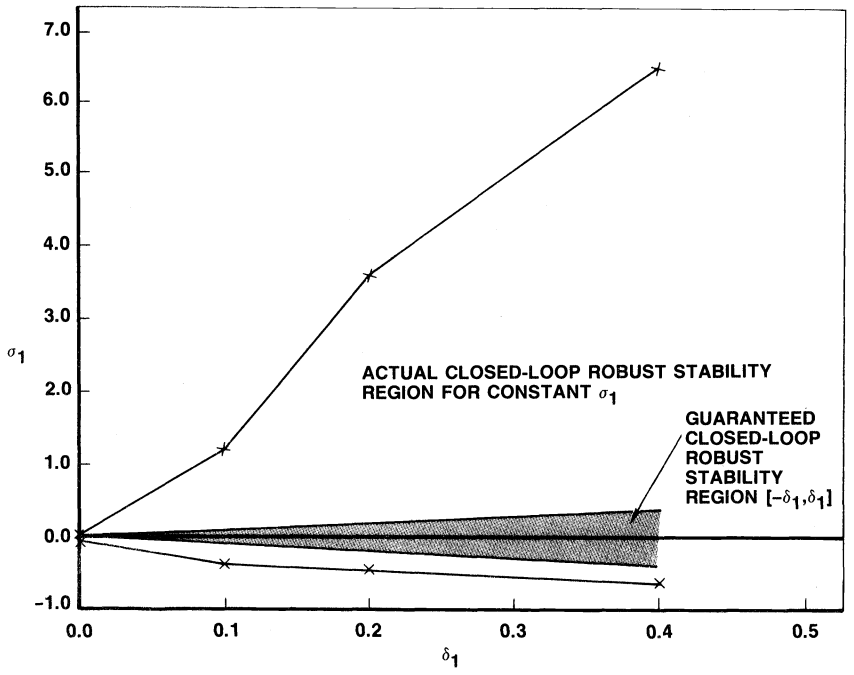

FIG. 1

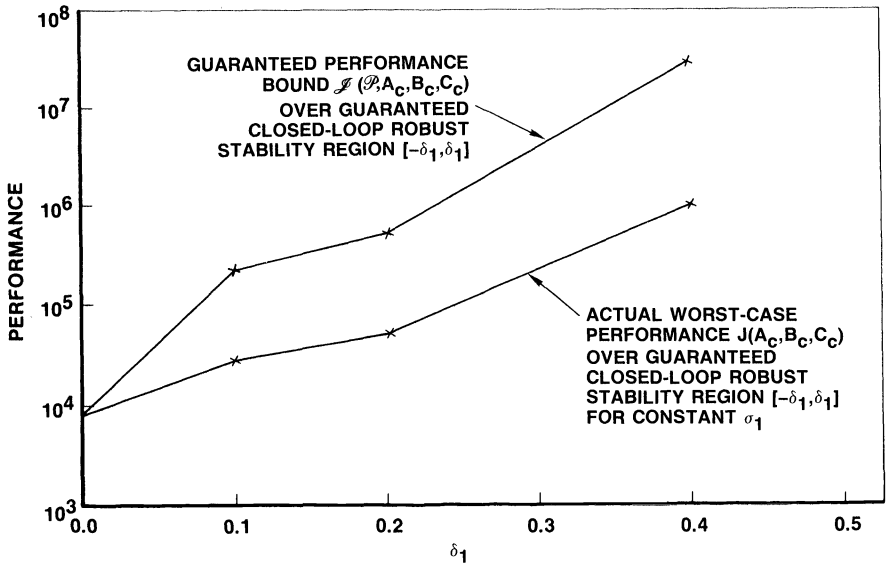

FIG. 2

TABLE 1

(δ_1, α_1)	A_c	B_c	C_c
(.1, 1)	$\begin{bmatrix} -14.917 & 1.0 \\ -85.177 & 3.9657 \end{bmatrix}$	$\begin{bmatrix} 15.917 \\ 79.959 \end{bmatrix}$	$[-5.2182 \quad -4.9657]$
(.2, 2)	$\begin{bmatrix} -17.963 & 1.0 \\ -133.65 & -4.4614 \end{bmatrix}$	$\begin{bmatrix} 18.963 \\ 127.05 \end{bmatrix}$	$[-6.6011 \quad -5.4614]$
(.4, 4)	$\begin{bmatrix} -47.813 & 1.0 \\ -1087.3 & -6.5463 \end{bmatrix}$	$\begin{bmatrix} 48.813 \\ 1073.5 \end{bmatrix}$	$[-13.766 \quad -7.5463]$

[38]-[40], controllers were obtained by solving (6.16)-(6.20) for $(\delta_1, \alpha_1) = (.1, 1), (.2, 2)$, and $(.4, 4)$. As stated previously, these numerical solutions also verify (4.2) with $\mathscr{P}$ given by (6.12). Figure 1 compares the guaranteed robust stability region to the "actual" robust stability region. This robust stability region was evaluated assuming constant $\hat{\sigma}_1(\cdot)$, although the theory actually guarantees robustness with respect to time-varying uncertainties. Thus, the gap between these regions may not be a reliable measure of the conservatism of the results. Note, however, that the design approach appears to provide more stability than is guaranteed a priori. This feature may be attributable to the desire for a symmetric stability interval so close to an unstabilizable plant perturbation, i.e., $\sigma_1 = -1$. Nevertheless, the stability design objectives have been met in accordance with Theorem 6.2. Interestingly, the form of the actual stability region mimics the classical 6-dB-downward/infinite-dB-upward gain margin of full-state-feedback LQR controllers [1]. Thus, this approach appears to provide an alternative to gain-margin recovery techniques [9], which address this specialized form of plant uncertainty. Finally, Fig. 2 compares guaranteed closed-loop performance to "actual" closed-loop performance over the guaranteed closed-loop robust stability region. Again the "actual" region was determined for constant $\hat{\sigma}_1(\cdot)$. Controller gains are given in

Table 1. Finally, we note that higher-order robust controllers were obtained for this example in [46] using the μ-function approach.

Acknowledgments. I thank Jill M. Straehla for preparing the manuscript versions of this paper, Scott W. Greeley for carrying out the numerical computations, Dr. Wassim M. Haddad for several helpful suggestions, and the reviewers for numerous helpful comments.

REFERENCES

[1] M. G. SAFONOV AND M. ATHANS, *Gain and phase margin for multiloop* LQG *regulators*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 173–179.

[2] J. C. DOYLE, *Guaranteed margins for* LQG *regulators*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 756–757.

[3] J. C. DOYLE AND G. STEIN, *Robustness with observers*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 607–611.

[4] ———, *Multivariable feedback design: concepts for a classical/modern synthesis*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 4–16.

[5] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301–320.

[6] J. C. DOYLE, *Analysis of feedback systems with structured uncertainties*, Proc. IEEE-D, 129 (1982), pp. 242–250.

[7] M. G. SAFONOV, *Stability margins of diagonally perturbed multivariable feedback systems*, Proc. IEEE-D, 129 (1982), pp. 251–256.

[8] G. ZAMES AND B. A. FRANCIS, *Feedback, minimax sensitivity, and optimal robustness*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 585–601.

[9] G. STEIN AND M. ATHANS, *The* LQG/LTR *procedure for multivariable feedback control design*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 105–114.

[10] B. A. FRANCIS, *A Course in H_∞ Control Theory*, Springer-Verlag, New York, 1987.

[11] E. SOROKA AND U. SHAKED, *On the robustness of* LQ *regulators*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 664–665.

[12] U. SHAKED AND E. SOROKA, *On the stability robustness of the continuous-time* LQG *optimal control*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1039–1043.

[13] M. TAHK AND J. L. SPEYER, *Modeling of parameter variations and asymptotic* LQG *synthesis*, in Proc. 24th IEEE Conference on Decision and Control, Athens, Greece, December 1986, pp. 1459–1465.

[14] S. S. L. CHANG AND T. K. C. PENG, *Adaptive guaranteed cost control of systems with uncertain parameters*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 474–483.

[15] R. V. PATEL, M. TODA, AND B. SRIDHAR, *Robustness of linear quadratic state feedback designs in the presence of system uncertainty*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 945–949.

[16] G. LEITMANN, *Guaranteed asymptotic stability for a class of uncertain linear dynamical systems*, J. Optim. Theory Appl., 27 (1979), pp. 96–106.

[17] A. VINKLER AND L. J. WOOD, *Multistep guaranteed cost control of linear systems with uncertain parameters*, J. Guid. Control, 2 (1979), pp. 449–456.

[18] M. ESLAMI AND D. L. RUSSELL, *On stability with large parameter variations: stemming from the direct method of Lyapunov*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1231–1234.

[19] M. CORLESS AND G. LEITMANN, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamical systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1139–1144.

[20] J. S. THORP AND B. R. BARMISH, *On guaranteed stability of uncertain linear systems via linear control*, J. Optim. Theory Appl., 35 (1981), pp. 559–579.

[21] B. R. BARMISH, M. CORLESS, AND G. LEITMANN, *A new class of stabilizing controllers for uncertain dynamical systems*, SIAM J. Control Optim., 21 (1983), pp. 246–255.

[22] B. R. BARMISH, I. R. PETERSEN, AND A. FEUER, *Linear ultimate boundedness control of uncertain dynamic systems*, Automatica, 19 (1983), pp. 523–532.

[23] B. R. BARMISH, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain linear system*, J. Optim. Theory Appl., 46 (1985), pp. 399–408.

[24] R. K. YEDAVALLI, S. S. BANDA, AND D. B. RIDGELY, *Time-domain stability robustness measures for linear regulators*, J. Guid. Control Dyn., 8 (1985), pp. 520–524.

[25] A. R. GALIMIDI AND B. R. BARMISH, *The constrained Lyapunov problem and its application to robust output feedback stabilization*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 410–419.

[26] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain systems*, Automatica, 22 (1986), pp. 397–411.

[27] 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.

[28] O. I. KOSMIDOU AND P. BERTRAND, *Robust-controller design for systems with large parameter variations*, Internat. J. Control, 45 (1987), pp. 927–938.

[29] D. C. HYLAND AND D. S. BERNSTEIN, *The optimal projection equations for fixed-order dynamic compensation*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1034–1037.

[30] 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., 23 (1986), pp. 122–151.

[31] D. S. BERNSTEIN, *The optimal projection equations for static and dynamic output feedback: the singular case*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1139–1143.

[32] D. S. BERNSTEIN AND W. M. HADDAD, LQG *control with an* H_∞ *performance bound: a Riccati equation approach*, IEEE Trans. Automat. Control, AC-34 (1989), to appear.

[33] D. S. BERNSTEIN, *Robust static and dynamic output feedback stabilization: deterministic and stochastic perspectives*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 1076–1084.

[34] Y. A. PHILLIS, *Controller design of systems with multiplicative noise*, IEEE Trans. Automat. Control, AC-30 (1985), pp. 1017–1019.

[35] D. S. BERNSTEIN AND S. W. GREELEY, *Robust controller synthesis using the maximum entropy design equations*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 362–364.

[36] D. S. BERNSTEIN AND D. C. HYLAND, *The optimal projection equations for reduced-order modelling, estimation and control of linear systems with multiplicative white noise*, J. Optim. Theory Appl., 58 (1988), pp. 387–409.

[37] S. RICHTER, *A homotopy algorithm for solving the optimal projection equations for fixed-order dynamic compensation: existence, convergence and global optimality*, in Proc. American Control Conference, Minneapolis, MN, June 1987, pp. 1527–1531.

[38] S. W. GREELEY AND D. C. HYLAND, *Reduced-order compensation*: LQG *reduction versus optimal projection using a homotopic continuation method*, Proc. IEEE Conference on Decision and Control, Los Angeles, CA, December 1987, pp. 742–747.

[39] A. GRUZEN, *Robust reduced order control of flexible structures*, Report CSDL-T-900, Charles Stark Draper Laboratory, Cambridge, MA, April 1986.

[40] A. GRUZEN AND W. E. VANDER VELDE, *Robust reduced-order control of flexible structures using the optimal projection/maximum entropy design methodology*, AIAA Guid. Nav. Contr. Conf., Williamsburg, VA, August 1978.

[41] J. W. BREWER, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 772–781.

[42] J. K. HALE, *Ordinary Differential Equations*, John Wiley, New York, 1969.

[43] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[44] L. ARNOLD, *Stochastic Differential Equations: Theory and Applications*, John Wiley, New York, 1974.

[45] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[46] A. PACKARD AND J. C. DOYLE, *Robust control with an* H_2 *performance objective*, in Proc. American Control Conference, Minneapolis, MN, June 1987, pp. 2141–2145.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 407-431, March 1989

009

ON THE STABILITY PROPERTIES OF SPLINE APPROXIMATIONS FOR RETARDED SYSTEMS*

F. KAPPEL† AND D. SALAMON‡

Abstract. This paper studies the qualitative properties of the spline approximation scheme for retarded functional differential equations introduced by Kappel and Salamon [*SIAM J. Control Optim.*, 25 (1987), pp. 1082-1117]. It is shown that the approximating systems are stable for large N if the underlying retarded functional differential equation is stable. In this case the approximating equations are in some sense uniformly (with respect to the approximation index) stable in the vector component of the state but not so in the complete state.

Key words. retarded functional differential equations, approximation, splines, controllability

AMS(MOS) subject classifications. 34K35, 41A15, 93D15

1. Introduction. In [10] and [11] we have introduced a new spline approximation scheme for retarded functional differential equations. The aim of this paper is to study the qualitative properties of this approximation scheme with particular emphasis on the stability problem.

The fundamental convergence properties of this approximation scheme have been established in [11]. The central result is a convergence proof for both the original semigroup $S(t)$ and its adjoint $S^*(t)$ in the strong operator topology. Here lies the main advantage over the spline approximation scheme, developed earlier in [2], for which the adjoint semigroup is only approximated in the weak operator topology. In addition, we have observed a quite significant improvement in the convergence behaviour of our numerical computations, some of which are reported in [11].

The main result of this paper is that the approximating systems (Σ^N) are stable (stabilizable, detectable) for sufficiently large N provided the original system (Σ) is stable (stabilizable and detectable.) The proof consists of three parts. The first part is a convenient characterization of the stability, stabilizability, and detectability of the approximating systems in terms of a certain characteristic matrix $\Delta^N(\lambda)$. The second part is a convergence proof for these matrices $\Delta^N(\lambda)$. The third part establishes a priori bounds for the unstable eigenvalues of the approximating systems.

We also discuss the role of the structural operator F in the spline approximation scheme. Moreover, we prove that the approximating systems cannot be stable in a uniform sense with respect to N and illustrate this result with computations of the spectrum. In this respect the spline approximation differs from the averaging approximation scheme in [1] for which the uniform exponential stability property has been established in [19]. But if we take the output of the system to be the vector component of the state, then the approximating systems are in a sense uniformly output stable with respect to N if the hereditary system is stable. For simplicity of presentation we restrict ourselves to the single delay case. All results are true for equations with multiple commensurate delays and without distributed delay. Some results are also true for the general case. For details see [10].

* Received by the editors December 23, 1986; accepted for publication (in revised form) May 13, 1988. This research was partly supported by Fonds zur Förderung der wissenschaftlichen Forschung (Austria) Project S3206.

† Institut für Mathematik, Karl-Franzens-Universität Graz, Elisabethstrasse 16, A-8010 Graz, Austria.

‡ Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom. The research of this author was partly supported by the National Science Foundation grant MCS-8210950.

2. Linear retarded control systems.

2.1. Functional differential equations. We consider the linear retarded functional differential equation (RFDE)

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-h) + B_0 u(t), \qquad y(t) = C_0 x(t), \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^l$, $y(t) \in \mathbb{R}^m$, A_0, $A_1 \in \mathbb{R}^{n\times n}$, $B_0 \in \mathbb{R}^{n\times l}$, $C_0 \in \mathbb{R}^{m\times n}$, and $h > 0$. It is obvious that (2.1) admits a unique solution $x(\cdot) \in L^2(-h, T; \mathbb{R}^n) \cap W^{1,2}(0, T; \mathbb{R}^n)$ for every input $u(\cdot) \in L^2(0, T; \mathbb{R}^l)$ and every initial condition of the form

$$x(0) = \phi^0, \quad x(\tau) = \phi^1(\tau), \quad -h \leqq \tau < 0, \tag{2.2}$$

where $\phi = (\phi^0, \phi^1) \in M^2 = \mathbb{R}^n \times L^2(-h, 0; \mathbb{R}^n)$ (see, for instance, [6], [9]). By $X(t) \in \mathbb{R}^{n\times n}$, $t \geqq -h$, we denote the fundamental matrix solution of (2.1), which corresponds to the initial condition $X(0) = I$, $X(\tau) = 0$, $-h \leqq \tau < 0$, and the input $u(t) \equiv 0$. Its Laplace transform is given by $\Delta(\lambda)^{-1}$, where $\Delta(\lambda) = \lambda I - A_0 - A_1 e^{-\lambda h}$, $\lambda \in \mathbb{C}$, is the characteristic matrix of (2.1).

2.2. State space theory. We consider two state concepts for (2.1). In the classical sense the state at time $t \geqq 0$ is defined to be the pair $z(t) = (x(t), x_t) \in M^2$, where $x_t(\tau) = x(t+\tau)$ for $-h \leqq \tau \leqq 0$. This state defines a weak solution of the abstract Cauchy problem

$$\dot{z}(t) = Az(t) + Bu(t), \quad z(0) = \phi, \quad y(t) = Cz(t), \tag{Σ}$$

where $B \in \mathscr{L}(\mathbb{R}^l, M^2)$ and $C \in \mathscr{L}(M^2, \mathbb{R}^m)$ are defined by $Bu = (B_0 u, 0)$ and $C\phi = C_0 \phi^0$ for $u \in \mathbb{R}^l$ and $\phi \in M^2$. The unbounded operator A: dom $A \to M^2$ is given by

$$A\phi = (A_0 \phi^1(0) + A_1 \phi^1(-h), \dot{\phi}^1), \qquad \text{dom } A = \{\phi \in M^2 \mid \phi^1 \in W^{1,2}, \phi^0 = \phi^1(0)\}$$

and generates a strongly continuous semigroup $S(t)$ of bounded linear operators on M^2. Therefore $z(t) \in M^2$ is given by the variation-of-constants formula

$$z(t) = S(t)\phi + \int_0^t S(t-s)Bu(s)\, ds. \tag{2.3}$$

Now let $S_T(t)$ denote the semigroup corresponding to the RFDE $\dot{x}(t) = A_0^T x(t) + A_1^T x(t-h)$ so that its generator A_T is defined as A with A_0, A_1 replaced by A_0^T, A_1^T. Then there is an alternative (dual) state concept for the RFDE (2.1) that relates the semigroups $S(t)$ and $S_T^*(t)$. It can be defined in terms of the structural operator $F \in \mathscr{L}(M^2)$ (for a normed linear space X we denote by $\mathscr{L}(X)$ the space of all bounded linear operators $X \to X$) given by

$$[F\phi]^0 = \phi^0, \quad [F\phi]^1(\sigma) = A_1 \phi(-h-\sigma), \quad -h \leqq \sigma \leqq 0 \tag{2.4}$$

for $\phi \in M^2$ (we define $\phi^1(\tau) = 0$ for $\tau \notin [-h, 0]$). It is a remarkable fact that for every weak solution $z(t) \in M^2$ of the Cauchy problem (Σ) the function $w(t) = Fz(t) \in M^2$ defines a weak solution of the abstract Cauchy problem

$$\dot{w}(t) = A_T^* w(t) + Bu(t), \quad w(0) = f \in M^2, \quad y(t) = Cw(t), \tag{Σ_T^*}$$

with $f = F\phi$.

Equivalently, the structural operator F satisfies the following equations:

$$FS(t) = S_T^*(t)F, \quad FB = B, \quad CF = C \tag{2.5}$$

for $t \geqq 0$. In particular, for every solution $x(t) \in \mathbb{R}^n$, $t \geqq -h$, of (2.1) the function $w(t) = F(x(t), x_t) \in M^2$ is given by

$$w(t) = S_T^*(t)F\phi + \int_0^t S_T^*(t-s)Bu(s)\, ds. \tag{2.6}$$

For more detailed discussion of these two state concepts and their relation see [6], [13], [16], and [18].

2.3. Stability, stabilizability, and controllability. System (2.1) is said to be *stable* if every solution $x(t)$ of the free system ($u(t) \equiv 0$) tends to zero as t goes to infinity. Equivalently, $\det \Delta(\lambda) = 0$ implies that $\operatorname{Re} \lambda < 0$ for $\lambda \in \mathbb{C}$ (see, for instance, [9]). Note that $\sigma(A) = \sigma(A_T^*) = \{\lambda \in \mathbb{C} \mid \det \Delta(\lambda) = 0\}$. Moreover, system (2.1) is said to be *stabilizable* if

$$\operatorname{rank} [\Delta(\lambda), B_0] = n \quad \text{for } \operatorname{Re} \lambda \geqq 0 \tag{2.7}$$

and *detectable* if

$$\operatorname{rank} \begin{bmatrix} \Delta(\lambda) \\ C_0 \end{bmatrix} = n \quad \text{for } \operatorname{Re} \lambda \geqq 0. \tag{2.8}$$

An abstract Cauchy problem is said to be *observable* if a nonzero initial state produces a nonzero output. Hence the Cauchy problem (Σ) is observable if and only if

$$y(t) = 0 \quad \text{for } t \geqq 0 \quad \text{implies } x(t) = 0 \quad \text{for all } t \geqq -h$$

for every solution of (2.1) and the Cauchy problem (Σ_T^*) is observable if and only if

$$y(t) = 0 \quad \text{for } t \geqq -h \quad \text{implies } x(t) = 0 \quad \text{for all } t \geqq -h$$

(see [17]). These two properties have been characterized as follows [13], [14], [17].

THEOREM 2.1. *System* (Σ) *is observable if and only if*

$$\operatorname{rank} \begin{bmatrix} \Delta(\lambda) \\ C_0 \end{bmatrix} = \operatorname{rank} A_1 = n \quad \textit{for all } \lambda \in \mathbb{C};$$

(Σ_T^*) *is observable if and only if*

$$\operatorname{rank} \begin{bmatrix} \Delta(\lambda) \\ C_0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} A_1 \\ C_0 \end{bmatrix} = n \quad \textit{for all } \lambda \in \mathbb{C}.$$

If (2.7) and (2.8) are satisfied, then there exist unique nonnegative, selfadjoint operators Π, $P \in L(M^2)$ satisfying range $\Pi \subset \operatorname{dom} A^*$, range $P \subset \operatorname{dom} A_T$, and the algebraic Riccati operator equations

$$A^*\Pi\phi = \Pi A\phi - \Pi BB^*\Pi\phi + C^*C\phi = 0, \tag{2.9}$$

$$A_T Pf + PA_T^* f - PBB^* Pf + C^*Cf = 0 \tag{2.10}$$

for $\phi \in \operatorname{dom} A$ and $f \in \operatorname{dom} A_T^*$ (see [4], [20]). It follows from (2.5) that the solution operators Π of (2.9) and P of (2.10) satisfy the identity

$$\Pi = F^*PF. \tag{2.11}$$

This was first observed in [5] for the Riccati differential equation. Finally, we point out that Π is injective if and only if (Σ) is observable, and that P is injective if and only if (Σ_T^*) is observable.

For a detailed discussion of the Riccati equation and its connection to optimal control theory see [4], [7], and [20].

3. Spline approximation.

3.1. Notation and terminology. Consider the finite-dimensional linear subspace

$$X^N = \left\{ \phi \in M^2 \;\middle|\; \phi^1 = \sum_{j=0}^{N} e_j^N z_j,\ z_j \in \mathbb{R}^n \right\},$$

where the scalar functions $e_j^N(\cdot) \in L^2(-h, 0)$ are defined by

$$e_0^N(\tau) = \begin{cases} \dfrac{N}{h}(\tau - t_1^N), & t_1^N \leqq \tau < t_0^N, \\ 0 \quad \text{elsewhere,} \end{cases}$$

$$e_j^N(\tau) = \begin{cases} \dfrac{N}{h}(\tau - t_{j+1}^N), & t_{j+1}^N \leqq \tau \leqq t_j^N, \\ -\dfrac{N}{h}(\tau - t_{j-1}^N), & t_j^N \leqq \tau \leqq t_{j-1}^N, \\ 0 \quad \text{elsewhere,} \end{cases}$$

$$e_N^N(\tau) = \begin{cases} -\dfrac{N}{h}(\tau - t_{N-1}^N), & t_N^N \leqq \tau \leqq t_{N-1}^N, \\ 0 \quad \text{elsewhere,} \end{cases}$$

for $j = 1, \cdots, N-1$ and meshpoints $t_j^N = -jh/N$ for $j = 0, \cdots, N$. Note that the function component of every $\phi \in X^N$ is a piecewise linear spline function on the interval $[-h, 0)$. The subspace $X^N \subset M^2$ can be identified with the Euclidean space $\mathbb{R}^{k(N)}$, $k(N) = n + (N+1)n$, via the embedding $\iota^N : \mathbb{R}^{k(N)} \to M^2$ defined by

$$\iota^N z = \left(z_0, \sum_{j=0}^{N} e_j^N z_j \right) \tag{3.1}$$

for $z = \operatorname{col}(z_0, z_1) \in \mathbb{R}^{k(N)}$, where $z_0 \in \mathbb{R}^n$ and $z_1 = \operatorname{col}(z_{10}, \cdots, z_{1N})$, $z_{ij} \in \mathbb{R}^n$, $j = 0, \cdots, N$. On $\mathbb{R}^{k(N)}$ we will always consider the induced inner product

$$\langle w, z \rangle_N = w^T Q^N z = \langle \iota^N w, \iota^N z \rangle_{M^2}, \tag{3.2}$$

where

$$Q^N = \begin{pmatrix} 1 & 0 \\ 0 & (h/N) q^N \end{pmatrix} \otimes I,$$

$$q^N = \begin{pmatrix} \frac{1}{3} & \frac{1}{6} & 0 & \cdots & 0 \\ \frac{1}{6} & \frac{2}{3} & & & 0 \\ 0 & & \ddots & & \vdots \\ & & & \frac{2}{3} & \frac{1}{6} \\ 0 & & \cdots & \frac{1}{6} & \frac{1}{3} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Here I denotes the $n \times n$ identity matrix. The corresponding vector and matrix norms will be denoted by $\| \cdot \|_N$. The adjoint operator $\pi^N = (\iota^N)^* : M^2 \to \mathbb{R}^{k(N)}$ is then given by

$$\pi^N \phi = (Q^N)^{-1} z, \qquad z_0 = \phi^0,$$

$$z_{1j} = \int_{-h}^{0} e_j^N(\tau) \phi^1(\tau)\, d\tau, \qquad j = 0, \cdots, N,$$

and satisfies the identities

$$\pi^N \iota^N = id, \qquad \iota^N \pi^N = p^N, \tag{3.3}$$

where $p^N: M^2 \to X^N$ denotes the orthogonal projection. Let the matrices $H^N \in \mathbb{R}^{k(N)\times k(N)}$, $B^N \in \mathbb{R}^{k(N)\times l}$, and $C^N \in \mathbb{R}^{m\times k(N)}$ be defined by

$$H^N = \left(\begin{array}{c|ccc} A_0 & 0 & \cdots & A_1 \\ \hline I & & & \\ 0 & & h^N \otimes I & \\ \vdots & & & \\ 0 & & & \end{array}\right), \qquad B^N = \begin{pmatrix} B_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$$C^N = (C_0 \quad 0 \quad \cdots \quad 0),$$

$$h^N = \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} & 0 & \cdots & 0 \\ \frac{1}{2} & 0 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & 0 & -\frac{1}{2} \\ & & & \frac{1}{2} & -\frac{1}{2} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Finally, we define

$$A^N = (Q^N)^{-1}H^N, \qquad A_T^N = (Q^N)^{-1}H_T^N,$$
$$(A_T^N)^* = (Q^N)^{-1}(H_T^N)^T, \qquad (A^N)^* = (Q^N)^{-1}(H^N)^T,$$

where the matrix $H_T^N \in \mathbb{R}^{(k(N)\times k(N))}$ is obtained from H^N by transposing the matrices A_0, A_1.

Now we consider two control systems on the state space $\mathbb{R}^{k(N)}$:

$$(\Sigma^N) \qquad \dot{z}^N(t) = A^N z^N(t) + B^N u(t), \quad z^N(0) = \pi^N \phi, \quad y^N(t) = C^N z^N(t),$$

$$(\Sigma_T^N{}^*) \qquad \dot{w}^N(t) = (A_T^N)^* w^N(t) + B^N u(t), \quad w^N(0) = \pi^N f, \quad y^N(t) = C^N w^N(t).$$

In [11] we establish the following convergence theorem.

THEOREM 3.1. (i) *For every $\phi \in M^2$ we have $\phi = \lim_{N\to\infty} p^N \phi$.*

(ii) *$B = \iota^N B^N$, $C = C^N \pi^N$ for every $N \in \mathbb{N}$.*

(iii) *There exist constants $M \geqq 1$, $\omega \geqq 0$ such that*

$$\|e^{A^N t}\|_N \leqq M e^{\omega t}, \qquad \|e^{(A_T^N)^* t}\|_N \leqq M e^{\omega t}$$

for every $t \geqq 0$ and every $N \in \mathbb{N}$.

(iv) *For all $\phi, f \in M^2$,*

$$S(t)\phi = \lim_{N\to\infty} \iota^N e^{A^N t} \pi^N \phi, \qquad S_T^*(t) f = \lim_{N\to\infty} \iota^N e^{(A_T^N)^* t} \pi^N f$$

and the limits are uniform on every compact time interval $[0, T]$.

In particular, this implies that for every $\phi \in M^2$ and every input $u(\cdot) \in L^2(0, T; \mathbb{R}^l)$ we have $z(t) = \lim_{N\to\infty} \iota^N z^N(t)$ for $0 \leqq t \leqq T$ (uniformly), where $z(t)$ is the unique weak solution of (Σ) and $z^N(t)$ satisfies (Σ^N). In the same manner the solutions $w^N(t)$ of $(\Sigma_T^N{}^*)$ approximate the solution $w(t)$ of (Σ_T^*).

In the remainder of this section we will study the structural properties of the approximating systems (Σ^N) and $(\Sigma_T^N{}^*)$.

3.2. The structural operator. In § 2 we have seen that the structural operator $F: M^2 \to M^2$ plays an important role for the state-space description of retarded systems. In this section we introduce an analogous operator for the description of the approximating systems (Σ^N) and (Σ_T^N). The first step in this direction is Lemma 3.2.

Suppose $u(\cdot)$ is an arbitrary control in $V_{(0,T)}[h]$ and its projection on the closed subspace $\bar{\mathcal{U}}$ is denoted by $u_p(\cdot)$. For each ψ in X one has

$$\begin{aligned}(u_p(\cdot), (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} &= (u(\cdot), (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} \\ &= (h, \psi)_X = (G\varphi, \psi)_X \\ &= ((\mathcal{S}_T B)^*(\cdot)\varphi, (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)},\end{aligned}$$

or

$$(u_p(\cdot) - (\mathcal{S}_T B)^*(\cdot)\varphi, (\mathcal{S}_T B)^*(\cdot)\psi)_{L^2(0,T;U)} = 0.$$

Since $\mathcal{U}$ is dense in $\bar{\mathcal{U}}$ and $[u_p(\cdot) - (\mathcal{S}_T B)^*(\cdot)\varphi] \in \mathcal{U}$, we have

$$u_p(\cdot) = (\mathcal{S}_T B)^*(\cdot)\varphi \quad \text{for each } u(\cdot) \in V_{(0,T)}[h].$$

Therefore

$$\|(\mathcal{S}_T B)^*(\cdot)\varphi\|_{L^2(0,T;U)} = \|u_p(\cdot)\|_{L^2(0,T;U)} \leqq \|u(\cdot)\|_{L^2(0,T;U)} \quad \text{for each } u(\cdot) \in V_{(0,T)}[h].$$

By uniqueness of the minimum norm optimal control of the linear system (1.5), (3.15) holds and $u^*(\cdot) \in \mathcal{U}$.

Since $u_\varepsilon = u^* - \varepsilon(\varepsilon + \hat{G})^{-1}u^*$ (see (3.10)), (3.14) is equivalent to

$$\lim_{\varepsilon \to 0} \varepsilon(\varepsilon + \hat{G})^{-1}u^* = 0 \quad \text{in } L_2(0, T; U). \tag{3.16}$$

If we consider another family with parameter $\varepsilon > 0$ of associated quadratic optimal control problems,

$$J_\varepsilon(v; \varphi) = \|(\mathcal{S}_T B)v - \varphi\|^2 + \varepsilon\|v(\cdot)\|^2_{L^2(0,T;U)},$$

then $J_\varepsilon(v; \varphi)$ takes its minimum at $v = v_\varepsilon$ defined by

$$v_\varepsilon = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\varphi = (\varepsilon + \hat{G})^{-1}u^*.$$

If we can show

$$\lim \varepsilon\|v_\varepsilon(\cdot)\|^2_{L^2(0,T;U)} = 0, \tag{3.17}$$

then

$$\begin{aligned}\lim_{\varepsilon \to 0} \varepsilon\|v_\varepsilon(\cdot)\|_{L^2(0,T;U)} &\leqq \lim_{\varepsilon \to 0} \max\{\varepsilon; \varepsilon\|v_\varepsilon(\cdot)\|_{L^2(0,T;U)}\} \\ &\leqq \lim_{\varepsilon \to 0} \max\{\varepsilon; \varepsilon\|v_\varepsilon(\cdot)\|^2_{L^2(0,T;U)}\} = 0.\end{aligned}$$

The last equation just is (3.16). The rest is to show (3.17) holds for $\varphi \in X$. (Notice, if $\varphi \in \bar{K}_{(0,T)}$ then (3.17) holds. Here we may show (3.17) holds for any given $\varphi \in X$.)

Suppose φ is arbitrarily given in X and

$$\varphi = \bar{\varphi} + \varphi^\perp,$$

where $\bar{\varphi} \in \bar{K}_{(0,T)}$ and $\varphi^\perp \in K^\perp_{(0,T)}$—the orthogonal complement of $K_{(0,T)}$. Since $X = \bar{K}_{(0,T)} \oplus K^\perp_{(0,T)}$, one has

$$\|\varphi\|^2 = \|\bar{\varphi}\|^2 + \|\varphi^\perp\|^2 \quad \text{for any } \varphi \in X.$$

Denote

$$\bar{v}_\varepsilon = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\bar{\varphi} \quad \text{and} \quad v^\perp_\varepsilon = (\varepsilon + \hat{G})^{-1}(\mathcal{S}_T B)^*\varphi^\perp.$$

Then

$$v_\varepsilon = \bar{v}_\varepsilon + v^\perp_\varepsilon.$$

3.3. Criteria for stability, stabilizability, and controllability. We shall need the following facts on the real $(N+1)\times(N+1)$-matrix $a^N=(q^N)^{-1}h^N$.

LEMMA 3.5. (a) *Let* $\|\cdot\|_N$ *be the operator norm corresponding to the vector norm* $|x|_N^2=\bar{x}^Tq^Nx$ *on* $\mathbb{C}^{N+1}$. *Then* $\|e^{a^Nt}\|_N\leqq 1$ *for* $N=1,2,\cdots$ *and* $t\geqq 0$.

(b) *Let* $\mu\in\sigma(a^N)$ *and* $x=\operatorname{col}(x_0,\cdots,x_N)\in\mathbb{C}^{N+1}$, $x\neq 0$, *such that either* $(\mu q^N-h^N)x=0$ *or* $(\mu q^N-(h^N)^T)x=0$. *Then* $x_0\neq 0$ *and* $x_N\neq 0$.

(c) *Let* $\mu\in\sigma(a^N)$ *and* $x=\operatorname{col}(1,0,\cdots,0)$ *or* $x=\operatorname{col}(0,\cdots,0,1)$. *Then* $x\notin\operatorname{range}(\mu q^N-h^N)$ *and* $x\notin\operatorname{range}(\mu q^N-(h^N)^T)$.

(d) $\operatorname{Re}\mu<0$ *for every* $\mu\in\sigma(a^N)$.

Proof. (a) For every $x\in\mathbb{C}^{N+1}$ the following equation holds:

$$\begin{aligned}\operatorname{Re}(\bar{x}^Th^Nx)&=(\operatorname{Re}x)^Th^N(\operatorname{Re}x)+(\operatorname{Im}x)^Th^N(\operatorname{Im}x)\\&=-\tfrac{1}{2}|x_0|^2-\tfrac{1}{2}|x_N|^2.\end{aligned}\tag{3.7}$$

Hence a^N is a dissipative operator on $\mathbb{C}^{N+1}$ with respect to the inner product

$$\langle y,x\rangle_N=\bar{y}^Tq^Nx,\qquad x,y\in\mathbb{C}^{N+1}.$$

Therefore $\exp(a^Nt)$, $t\geqq 0$, is a contraction semigroup on $\mathbb{C}^{N+1}$ supplied with the norm $|\cdot|_N$ (see, for instance, [15]).

(b) This follows from

$$\mu q^N-h^N=\begin{pmatrix}\frac{\mu}{3}+\frac{1}{2} & \frac{\mu}{6}+\frac{1}{2} & 0 & \cdots & 0\\ \frac{\mu}{6}-\frac{1}{2} & \frac{2\mu}{3} & & & \\ & & & & 0\\ & & & & \\ 0 & & & & \frac{\mu}{6}+\frac{1}{2}\\ & & & & \frac{2\mu}{3}\\ 0 & \cdots & 0 & \frac{\mu}{6}-\frac{1}{2} & \frac{\mu}{3}+\frac{1}{2}\end{pmatrix}$$

and the fact that $\mu=\pm 3$ is not an eigenvalue of a^N.

(c) $x\in\operatorname{range}(\mu q^N-h^N)$ would imply $x\perp\ker(\mu q^N-(h^N)^T)$, which is impossible by (b).

(d) Assume that $\mu\in\sigma(a^N)$ and $\operatorname{Re}\mu\geqq 0$. Then there exists an $x\in\mathbb{C}^{N+1}$, $x\neq 0$, such that $(\mu q^N-h^N)x=0$. By (3.7) this implies

$$\begin{aligned}0\leqq(\operatorname{Re}\mu)\bar{x}^Tq^Nx&=\operatorname{Re}(\bar{x}^Th^Nx)\\&=-\tfrac{1}{2}|x_0|^2-\tfrac{1}{2}|x_N|^2.\end{aligned}$$

Hence $x_0=x_N=0$, and therefore $x=0$ by (b) in contradiction to $x\neq 0$. □

For every $\mu\in\mathbb{C}$ not in the spectrum of a^N (in particular, for every μ in the closed right half-plane) we introduce the vector

$$\alpha^N(\mu)=\operatorname{col}(\alpha_0^N(\mu),\cdots,\alpha_N^N(\mu))\tag{3.8}$$

as the unique solution of

$$(\mu q^N-h^N)\alpha^N(\mu)=\operatorname{col}(1,0,\cdots,0).\tag{3.9}$$

The complex $n\times n$-matrix

$$\Delta^N(\lambda)=\lambda I-A_0-A_1\alpha_N^N\left(\frac{\lambda h}{N}\right)\tag{3.10}$$

plays a role for the approximating systems (Σ^N) and (Σ_T^{N*}) analogous to that of the characteristic matrix $\Delta(\lambda)$ for the original systems (Σ) and (Σ_T^*). In particular, it determines their input-output behaviour (see Proposition 3.7 below).

Next we characterize the resolvent operator $(\lambda Q^N - H^N)^{-1}$ in terms of the matrices $\Delta^N(\lambda)$, F^N and $E^N(\lambda) \in \mathbb{C}^{k(N)\times n}$, $T^N(\lambda) \in \mathbb{C}^{k(N)\times k(N)}$. The latter are defined as follows:

$$E^N(\lambda) = \operatorname{col}\left(1, \alpha_0^N\left(\frac{\lambda h}{N}\right), \cdots, \alpha_N^N\left(\frac{\lambda h}{N}\right)\right) \otimes I, \tag{3.11}$$

$$T^N(\lambda) = \begin{pmatrix} 0 & 0 \\ 0 & ((\lambda h/N)q^N - h^N)^{-1} \end{pmatrix} \otimes I. \tag{3.12}$$

PROPOSITION 3.6. *Let $\lambda \in \mathbb{C}$ with $\lambda h/N \notin \sigma(a^N)$ and $x, z \in \mathbb{C}^{k(N)}$, where x is written as $x = (x_0, x_1)$ with $x_0 \in \mathbb{R}^n$, then*

(a) $(\lambda Q^N - H^N)x = z$ *if and only if*

$$x = E^N(\lambda)x_0 + T^N(\lambda)z, \tag{3.13.1}$$

$$\Delta^N(\lambda)x_0 = E^N(\lambda)^T F^N z. \tag{3.13.2}$$

(b) $(\lambda Q^N - (H_T^N)^T)x = z$ *if and only if*

$$x = F^N E^N(\lambda)x_0 + T^N(\lambda)^T z, \tag{3.14.1}$$

$$\Delta^N(\lambda)x_0 = E^N(\lambda)^T z. \tag{3.14.2}$$

Proof. We put $z = (z_0, z_1)$, $x_1 = (x_{10}, \cdots, x_{1N})$, $z_1 = (z_{10}, \cdots, z_{1N})$. It is easy to see that $(\lambda Q^N - H^N)x = z$ if and only if

$$(\lambda I - A_0)x_0 - A_1 x_{1N} = z_0, \tag{3.15.1}$$

$$\left(\left(\frac{\lambda h}{N} q^N - h^N\right) \otimes I\right) x_1 = z_1 + \operatorname{col}(x_0, 0, \cdots, 0). \tag{3.15.2}$$

Observing that (see (3.9))

$$\left(\frac{\lambda h}{N} q^N - h^N\right)^{-1} = \begin{pmatrix} \alpha_0^N\left(\frac{\lambda h}{N}\right) & * \text{---} * \\ \vdots & | \quad\quad | \\ \alpha_N^N\left(\frac{\lambda h}{N}\right) & \cdots \quad \alpha_0^N\left(\frac{\lambda h}{N}\right) \end{pmatrix},$$

we get from (3.15.2):

$$x_1 = \left(\left(\frac{\lambda h}{N} q^N - h^N\right) \otimes I\right)^{-1} z_1 + E^N(\lambda)x_0,$$

which proves (3.13.1) and

$$x_{1N} = \alpha_N^N\left(\frac{\lambda h}{N}\right)x_0 + \sum_{k=0}^{N} \alpha_{N-k}^N\left(\frac{\lambda h}{N}\right) z_{1k}.$$

The last expression together with (3.15.1) establishes (3.13.2).

For (b) we observe that $(\lambda Q^N - (H_T^N)^T)x = z$ is equivalent to

$$\lambda x_0 - A_0 x_0 - x_{10} = z_0, \tag{3.16.1}$$

$$\left(\left(\frac{\lambda h}{N} q^N - (h^N)^T\right) \otimes I\right) x_1 = z_1 + \operatorname{col}(0, \cdots, 0, A_1 x_0). \tag{3.16.2}$$

Observing

$$\left(\frac{\lambda h}{N} q^N - (h_T^N)^T\right)^{-1} = \begin{pmatrix} \alpha_0^N\left(\frac{\lambda h}{N}\right) & \cdots & \alpha_N^N\left(\frac{\lambda h}{N}\right) \\ * & & \vdots \\ \vdots & & \\ * & & \alpha_0^N\left(\frac{\lambda h}{N}\right) \end{pmatrix},$$

we get from (3.16.2):

$$x = F^N E^N(\lambda) x_0 + T^N(\lambda) z,$$

which proves (3.14.1), and

$$x_{10} = \sum_{k=0}^{N} \alpha_k^N\left(\frac{\lambda h}{N}\right) z_{1k} + A_1 \alpha_N^N\left(\frac{\lambda h}{N}\right) x_0,$$

which together with (3.16.1) gives (3.14.2). □

In particular, the previous proposition shows that

$$(\lambda Q^N - H^N)^{-1} = E^N(\lambda)\Delta^N(\lambda)^{-1} E^N(\lambda)^T F^N + T^N(\lambda),$$
$$(\lambda Q^N - (H_T^N)^T)^{-1} = F^N E^N(\lambda)\Delta^N(\lambda)^{-1} E^N(\lambda)^T + T^N(\lambda)^T,$$

and hence

$$(\lambda I - A^N)^{-1} = E^N(\lambda)\Delta^N(\lambda)^{-1} E^N(\lambda)^T Q^N F^N + T^N(\lambda) Q^N, \tag{3.17}$$

$$(\lambda I - (A_T^N)^*)^{-1} = F^N E^N(\lambda)\Delta^N(\lambda)^{-1} E^N(\lambda)^T Q^N + T^N(\lambda)^T Q^N \tag{3.18}$$

provided $\lambda \notin \sigma(a^N)$ and $\det \Delta^N(\lambda) \neq 0$.

PROPOSITION 3.7. (a) *The left upper $n \times n$ block $X^N(t)$ in the matrix $e^{A^N t}$ coincides with that of the matrix $e^{(A_T^N)^* t}$. Its Laplace transform is given by $\Delta^N(\lambda)^{-1}$ (see* (3.10) *for the definition of $\Delta^N(\lambda)$).*

(b) *Let $w^N(t) = \operatorname{col}(w_0^N(t), \cdots)$ and $z^N(t) = \operatorname{col}(z_0^N(t), \cdots)$ be the unique solutions of (Σ^N) and $(\Sigma_T^N{}^*)$, respectively, with initial state zero. Then*

$$w_0^N(t) = z_0^N(t) = \int_0^t X^N(t-s) B_0 u(s)\, ds, \qquad t \geqq 0.$$

(c) *The transfer matrices of (Σ^N) and $(\Sigma_T^N{}^*)$ coincide and are given by*

$$G^N(\lambda) = C_0 \Delta^N(\lambda)^{-1} B_0.$$

Proof. Statement (a) is an immediate consequence of (3.17), (3.18), and the special form of the matrices $E^N(\lambda)$, $T^N(\lambda)$, f^N, Q^N. Statements (b) and (c) follow directly from (a). □

The following characterization of stabilizability and detectability for the approximating systems (Σ^N) and $(\Sigma_T^N{}^*)$ is precisely the analogue to (2.7) and (2.8).

THEOREM 3.8. (a) *For $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ the following properties are equivalent:*

(i) $\lambda \in \sigma(A^N)$;

(ii) $\lambda \in \sigma((A_T^N)^*)$;

(iii) $\det \Delta^N(\lambda) = 0$.

In particular, the matrix A^N (or equivalently $(A_T^N)^*$) is stable if and only if $\det \Delta^N(\lambda) \neq 0$ for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$.

(b) *The system (Σ^N) (or equivalently $(\Sigma_T^N{}^*)$) is stabilizable if and only if*

$$\operatorname{rank}[\Delta^N(\lambda), B_0] = n$$

for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$.

(c) *The system* (Σ^N) *(or equivalently* (Σ_T^N*)*) is detectable if and only if*

$$\operatorname{rank}\begin{pmatrix}\Delta^N(\lambda)\\ C_0\end{pmatrix}=n$$

for every $\lambda\in\mathbb{C}$ *with* $\operatorname{Re}\lambda\geqq 0$.

Proof. It is well known from finite-dimensional linear system theory that (Σ^N) is detectable if and only if $\ker(\lambda I-A^N)\cap\ker C^N=\{0\}$, or equivalently,

$$\ker(\lambda Q^N-H^N)\cap\ker C^N=\{0\}$$

for every $\lambda\in\mathbb{C}$ with $\operatorname{Re}\lambda\geqq 0$. But $\lambda(h/N)\notin\sigma(a^N)$ for $\operatorname{Re}\lambda\geqq 0$ (Lemma 3.5(d)). Therefore, according to Proposition 3.6(a), $x=\operatorname{col}(x_0,x_1)\in\ker(\lambda Q-H^N)$ is equivalent to $\Delta^N(\lambda)x_0=0$ and $x=E^N(\lambda)x_0$. This implies that detectability of (Σ^N) is equivalent to

$$\ker\Delta^N(\lambda)\cap\ker C^N=\{0\}$$

for every $\lambda\in\mathbb{C}$ with $\operatorname{Re}\lambda\geqq 0$. When we use Proposition 3.6(b), it follows analogously that this condition is also equivalent to detectability of (Σ_T^N*).

Statement (b) follows from (c) by duality and statement (a) follows from (b) with $B_0=0$. □

THEOREM 3.9. (a) *Let* $\lambda\in\mathbb{C}$ *be such that* $\lambda(h/N)\notin\sigma(a^N)$. *Then* $\lambda\in\sigma(A^N)$ *if and only if* $\det\Delta^N(\lambda)=0$. *If* $\lambda(h/N)\in\sigma(a^N)$, *then* $\lambda\in\sigma(A^N)$ *if and only if* $\det A_1=0$. *Moreover,* $\sigma((A_T^N)*)=\sigma(A^N)$.

(b) *System* (Σ^N) *is controllable if and only if*

$$\operatorname{rank}[\Delta^N(\lambda),B_0]=n\quad\text{for all }\lambda\in\mathbb{C}\backslash\sigma\left(\frac{N}{h}a^N\right),$$

$$\operatorname{rank}[A_1,b_0]=n.$$

(c) *System* (Σ^N) *is observable if and only if*

$$\operatorname{rank}\begin{pmatrix}\Delta^N(\lambda)\\ C_0\end{pmatrix}=n\quad\text{for all }\lambda\in\mathbb{C}\backslash\sigma\left(\frac{N}{h}a^N\right),$$

$$\operatorname{rank}A_1=n.$$

(d) *System* (Σ_T^N*) *is controllable if and only if*

$$\operatorname{rank}[\Delta^N(\lambda),B_0]=n\quad\text{for all }\lambda\in\mathbb{C}\backslash\sigma\left(\frac{N}{h}a^N\right),$$

$$\operatorname{rank}A_1=n.$$

(e) *System* (Σ_T^N*) *is observable if and only if*

$$\operatorname{rank}\begin{pmatrix}\Delta^N(\lambda)\\ C_0\end{pmatrix}=n\quad\text{for all }\lambda\in\mathbb{C}\backslash\sigma\left(\frac{N}{h}a^N\right),$$

$$\operatorname{rank}\begin{pmatrix}A_1\\ C_0\end{pmatrix}=n.$$

Proof. We first prove (c), i.e., we must show that $\ker(\lambda I-A^N)\cap\ker C^N=\{0\}$ for all $\lambda\in\mathbb{C}$. For $\lambda\notin\sigma((N/h)a^N)$, or equivalently, for $\lambda h/N\notin\sigma(a^N)$ we see as in the proof of Theorem 3.8(c) that this is equivalent to $\operatorname{rank}\binom{\Delta^N(\lambda)}{C_0}=n$.

Now let $\lambda \in \sigma((N/h)a^N)$. We assume $\operatorname{rank} A_1 = n$ and take $x = \operatorname{col}(x_0, x_{10}, \cdots, x_{1N}) \in \ker(\lambda I - A^N) \cap \ker C^N$. Then (3.15.2) implies $(((\lambda h/N)q^N - h^N) \otimes I)x_1 = \operatorname{col}(x_0, 0, \cdots, 0)$ and therefore $x_0 = 0$ by Lemma 3.5(c). This and (3.15.1) imply $A_1 x_{1N} = 0$, i.e., $x_{1N} = 0$. Then we get from Lemma 3.5(b) $x_1 = 0$, and thus $x = 0$.

Conversely assume that $\ker(\lambda I - A^N) \cap \ker C^N = \{0\}$ and take $\xi \in \ker A_1$. According to Lemma 3.5(b) there exists a vector $x_1 = \operatorname{col}(x_{10}, \cdots, x_{1N}) \in \mathbb{C}^{(N+1)n}$ such that $[(\lambda(h/N)q^N - h^N) \otimes I]x_1 = 0$ and $x_{1N} = \xi$. For $x = \operatorname{col}(0, x_1)$ equations (3.15) (with $z_0 = 0$) imply $(\lambda Q^N - H^N)x = 0$. Obviously we have $C^N x = 0$. By assumption this implies $x = 0$, and thus $\xi = 0$. We conclude $\operatorname{rank} A_1 = n$. This finishes the proof of statement (c).

It still holds that $\lambda \in \sigma((N/h)a^N)$. Assume that $\ker A_1 \cap \ker c_0 = \{0\}$ and let $x \in \mathbb{C}^{k(N)}$ satisfy $(\lambda Q - (H_T^N)^T)x = 0$, $C^N x = 0$. Then (3.16.2) and Lemma 3.5(c) imply $A_1 x_0 = 0$. This together with $C_0 x_0 = 0$ shows $x_0 = 0$. Hence it follows by (3.16.1) that $x_{10} = 0$. Finally we get from Lemma 3.5(b) that $x_1 = 0$, and thus $x = 0$, i.e., $\ker(\lambda I - (A^N)^*) \cap \ker C^N = \{0\}$.

Conversely, suppose that $\ker(\lambda I - (A_T^N)^*) \cap \ker C^N = \{0\}$ and let $x_0 \in \ker A_1 \cap \ker C_0$. By Lemma 3.5(b) there exists a vector $\alpha = \operatorname{col}(\alpha_0, \cdots, \alpha_N) \in \mathbb{C}^{N+1}$ such that

$$\left(\lambda \frac{h}{N} q^N - (h^N)^T\right)\alpha = 0 \quad \text{and} \quad \alpha_0 = 1.$$

We define $x_1 = \operatorname{col}(x_{10}, \cdots, x_{1N}) \in \mathbb{C}^{n(N+1)}$ by

$$x_{1k} = \alpha_k(\lambda I - A_0)x_0$$

for $k = 0, \cdots, N$. Then it follows from (3.16) that $x = \operatorname{col}(x_0, x_1) \in \ker(\lambda Q^N - (H_T^N)^T) \cap \ker C^N$. By assumption this implies $x = 0$, and hence $x_0 = 0$, i.e., $\ker A_1 \cap \ker C_0 = \{0\}$. Thus statement (e) is proved.

Statements (b) and (d) follow from (e) and (c) by duality. The proof of statement (a) is the same as for (c) with $C^N = 0$, respectively $C_0 = 0$. □

4. Stability. It is our goal to prove that stability (respectively, stabilizability, or detectability) of the original system (Σ) implies the corresponding property for the approximating system (Σ^N) and (Σ_T^{N*}) provided N is sufficiently large. The first step in this direction was the characterization of these properties in Theorem 3.8 using the matrices $\Delta^N(\lambda)$. The second step will be a convergence result for the characteristic matrices $\Delta^N(\lambda)$. As a third step we need a priori bounds for the unstable eigenvalues of the matrices A^N.

4.1. Convergence of $\Delta^N(\lambda)$. First we derive explicit formulas for the $\alpha_k^N(\mu)$ (as defined in (3.8) and (3.9)) and use those to prove convergence of $\Delta^N(\lambda)$ to $\Delta(\lambda)$. Let the rational functions $d_k^N(\mu)$, $k = 0, \cdots, N$, and the polynomials $p_k(\mu)$, $q_k(\mu)$, $k = -1, \cdots, N$, respectively, $k = 1, \cdots, N$, be defined recursively by

$$\begin{aligned} &d_0^N(\mu) = 2\mu + 3, \\ &d_k^N(\mu) = 4\mu + \frac{9 - \mu^2}{d_{k-1}^N(\mu)}, \qquad k = 1, \cdots, N-1, \\ &d_N^N(\mu) = 2\mu + 3 + \frac{9 - \mu^2}{d_{N-1}^N(\mu)}, \end{aligned} \tag{4.1}$$

$$\begin{aligned} &p_{-1}(\mu) = 1, \qquad p_0(\mu) = 2\mu + 3, \\ &p_k(\mu) = 4\mu p_{k-1}(\mu) + (9 - \mu^2)p_{k-2}(\mu), \\ &q_k(\mu) = (2\mu + 3)p_{k-1}(\mu) + (9 - \mu^2)p_{k-2}(\mu). \end{aligned} \tag{4.2}$$

The function $w = w(\mu)$ is defined by

$$w(\mu) = (9+3\mu^2)^{1/2}, \qquad \mu \in \mathbb{C}, \tag{4.3}$$

taking the branch that is positive for $\mu = i\theta$, $\theta \in \mathbb{R}$, $|\theta| < \sqrt{3}$. Furthermore, we set

$$\gamma_0(\mu) = 2\mu + w(\mu), \quad \gamma_1(\mu) = 2\mu - w(\mu), \quad \mu \in \mathbb{C}. \tag{4.4}$$

Since $w(\mu)$ is real for $\mu = i\theta$, $|\theta| \leqq \sqrt{3}$, we obtain

$$|\gamma_0(i\theta)| = |\gamma_1(i\theta)| = |3 - i\theta|, \quad |\theta| \leqq \sqrt{3}, \quad \theta \in \mathbb{R}. \tag{4.5}$$

Therefore the function $\delta = \delta(\theta)$ is well defined by

$$e^{i\delta(\theta)} = \frac{\gamma_1(i\theta)}{\gamma_0(i\theta)} = \frac{7\theta^2 - 9 + 4i\theta w(i\theta)}{9+\theta^2}, \quad |\theta| \leqq \sqrt{3}, \quad \theta \in \mathbb{R},$$
$$0 \leqq \delta(\theta) \leqq 2\pi. \tag{4.6}$$

LEMMA 4.1. (a) *For $k = 0, \cdots, N$ and $\mu \notin \sigma(a^N)$,*

$$\alpha_k^N(\mu) = \frac{6(3-\mu)^k}{d_{N-k}^N(\mu) \cdot \cdots \cdot d_N^N(\mu)} = \frac{6(3-\mu)^k p_{N-k-1}(\mu)}{q_N(\mu)}. \tag{4.7}$$

(b) $\det(\mu q^N - h^N) = (1/6^{N+1}) q_N(\mu)$ *and*

$$2 \cdot 6^{N+1} w \det(\mu q^N - h^N) = (3+w)^2(\gamma_0)^N - (3-w)^2(\gamma_1)^N$$

for all $\mu \in \mathbb{C}$.

(c) *For $\mu \notin \sigma(a^N)$ and $\mu \neq \pm i\sqrt{3}$,*

$$\alpha_k^N(\mu) = 6(3-\mu)^k \frac{(3+w)(\gamma_0)^{N-k} - (3-w)(\gamma_1)^{N-k}}{(3+w)^2(\gamma_0)^N - (3-w)^2(\gamma_1)^N}, \tag{4.8}$$

$k = 0, \cdots, N$. Moreover,

$$|\alpha_{N-k}^N(i\theta)|^2 = 36 \frac{9(1-\cos k\delta(\theta)) + w^2(i\theta)(1+\cos k\delta(\theta))}{(9+w^2(i\theta))^2(1-\cos N\delta(\theta)) + 36w^2(i\theta)(1+\cos N\delta(\theta))} \tag{4.9}$$

for $k = 0, \cdots, N$, $|\theta| < \sqrt{3}$, $\theta \in \mathbb{R}$.

Proof. Suppose that the functions $d_k = d_k^N(\mu)$, $k = 0, \cdots, N$, are given by (4.1) and define

$$b_k = \frac{-3-\mu}{d_{k-1}}, \quad c_k = \frac{3-\mu}{d_{k-1}}, \quad k = 1, \cdots, N.$$

Then it is easy to see that

$$6(\mu q^N - h^N) = \begin{pmatrix} 1 & -b_N & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & -b_1 \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} d_N & 0 & \cdots & 0 \\ 0 & \ddots & & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & d_0 \end{pmatrix}$$
$$\times \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -c_N & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

It is not difficult to calculate the inverse matrices. Since $\alpha^N(\mu)$ is the first column of $(\mu q^N - h^N)^{-1}$ (see (3.9)) we conclude that

$$\alpha_k^N(\mu) = \frac{6c_{N-k+1} \cdot \cdots \cdot c_N}{d_N} = \frac{6(3-\mu)^k}{d_{N-k} \cdot \cdots \cdot d_N},$$

$k = 0, \cdots, N$.

If the polynomials $p_k(\mu)$ and $q_k(\mu)$ are given by (4.2) and the rational functions $d_k^N(k)$ by (4.1), then we see by induction that

$$d_k^N(\mu) = p_k(\mu)/p_{k-1}(\mu), \qquad k = 0, \cdots, N-1,$$
$$d_N^N(\mu) = q_N(\mu)/p_{N-1}(\mu).$$

This implies

$$p_k(\mu) = d_0^N(\mu) \cdot \cdots \cdot d_k^N(\mu), \qquad k = 0, \cdots, N-1,$$
$$q_N(\mu) = d_0^N(\mu) \cdot \cdots \cdot d_N^N(\mu) = \det[6(\mu q^N - h^N)].$$

This finishes the proof of (a) and also establishes the first part of (b).

To prove (c) we choose $\mu \in \mathbb{C}$, $\mu \neq \pm i\sqrt{3}$. Then $\gamma_0 \neq \gamma_1$ and

$$\gamma_i^2 - 4\mu\gamma_i - (9-\mu^2) = 0, \qquad i = 0, 1.$$

Hence γ_0 and γ_1 are the characteristic roots of the difference equation in (4.2). This implies that

$$p_k(\mu) = \frac{3+w}{2w}(\gamma_0)^{k+1} - \frac{3-w}{2w}(\gamma_1)^{k+1}, \tag{4.10}$$

$k = -1, 0, \cdots, N-1$. Using $\gamma_0\gamma_1 = \mu^2 - 9$, we get from (4.2) and (4.10) that

$$\begin{aligned} q_N(\mu) &= (2\mu+3)p_{N-1}(\mu) + (9-\mu^2)p_{N-2}(\mu) \\ &= \frac{3+w}{2w}[(2\mu+3)(\gamma_0)^N + (9-\mu^2)(\gamma_0)^{N-1}] \\ &\quad - \frac{3+w}{2w}[(2\mu+3)(\gamma_1)^N + (9-\mu^2)(\gamma_1)^{N-1}] \\ &= \frac{(3+w)^2}{2w}(\gamma_0)^N - \frac{(3-w)^2}{2w}(\gamma_1)^N. \end{aligned} \tag{4.11}$$

The second part of (b) and (4.8) are immediate consequences of (4.10), (4.11).

To prove (4.9) we use (4.8) and observe $\gamma_1(i\theta) = e^{i\delta(\theta)}\gamma_0(i\theta)$ and (4.5). □

The explicit formulas in the previous lemma allow us to prove that the matrices $\Delta^N(\lambda)$ actually converge to the characteristic matrix $\Delta(\lambda)$ of the delay system.

THEOREM 4.2. $\Delta(\lambda) = \lim_{N\to\infty} \Delta^N(\lambda)$, $\lambda \in \mathbb{C}$, *the limit being uniform on bounded subsets of* $\mathbb{C}$.

Proof. Fix $\delta \in (0, \sqrt{3})$. Then $w(\mu)$ as defined in (4.3) is continuous and $|w(\mu)+3| \geqq 3$ on $|\mu| \leqq \delta$. From $w(\mu) - 3 = 3\mu^2/(w(\mu)+3)$ we see that

$$|w(\mu) - 3| \leqq |\mu|^2 \quad \text{if } |\mu| \leqq \delta. \tag{4.12}$$

In the next step we prove that $\alpha_N^N(\mu/N)$ converges for arbitrary $c > 0$ uniformly to $e^{-\mu}$ on $|\mu| \leqq c$ as $N \to \infty$. To this end we use formula (4.8) for $k = N$ and obtain, with $w = w(\mu/N)$, $N \geqq c\delta^{-1}$,

$$\alpha_N^N\left(\frac{\mu}{N}\right)^{-1} = \frac{(3+w)^2}{12w}\left(w + 2\frac{\mu}{N}\Big/3 - \frac{\mu}{N}\right)^N - \frac{(3-w)^2}{12w}\left(-w - 2\frac{\mu}{N}\Big/3 - \frac{\mu}{N}\right)^N. \tag{4.13}$$

From (4.12) and $\lim_{N\to\infty} w(\mu/N) = 3$ uniformly on $|\mu| \leqq c$ we see that

$$\lim_{N\to\infty} \frac{(3+w(\mu/N))^2}{12w(\mu/N)} = 1 \quad \text{and} \quad \lim_{N\to\infty} \frac{(3-w(\mu/N))^2}{12w(\mu/N)} = 0$$

uniformly on $|\mu| \leqq c$. Moreover, we also obtain from (4.12):

$$\frac{w+2\mu/N}{3-\mu/N} = 1+\frac{\mu}{N}+\frac{w-3+(\mu/N)^2}{3-\mu/N} = 1+\frac{\mu}{N}+O\left(\frac{1}{N^2}\right),$$

$$\frac{w-2\mu/N}{3-\mu/N} = 1-\frac{\mu}{3N}+\frac{w-3-\frac{1}{3}(\mu/N)^2}{3-\mu/N} = 1-\frac{\mu}{3N}+O\left(\frac{1}{N^2}\right)$$

as $N \to \infty$ uniformly on $|\mu| \leqq c$. These relations together with (4.13) show

$$\lim_{N\to\infty} \alpha_N^N\left(\frac{\mu}{N}\right)^{-1} = e^{\mu}$$

uniformly on $|\mu| \leqq c$. Finally, the theorem follows from (3.10). □

4.2. Uniform bounds. We first establish bounds for the $\alpha_N^N(\mu)$ in $\operatorname{Re}\mu \geqq 0$.

LEMMA 4.3. *The estimate* $|\alpha_N^N(\mu)| \leqq 2$ *is valid for all* $\mu \in \mathbb{C}$ *with* $\operatorname{Re}\mu \geqq 0$ *and all* $N = 1, 2, \cdots$.

Proof. Since, according to Lemmas 3.5(d) and 4.1(b), the polynomial $q_N(\mu)$ is stable, $\alpha_N^N(\mu)$ is a proper rational function without poles in $\operatorname{Re}\mu \geqq 0$ (cf. (4.7)). It follows from the maximum principle for analytic functions that $|\alpha_N^N(\mu)|$ achieves its maximum value in $\operatorname{Re}\mu \geqq 0$ on the imaginary axis. Therefore we only have to prove $|\alpha_N^N(i\omega)| \leqq 2$ for all $\omega \in \mathbb{R}$ and all N.

First we consider $\mu \in i\mathbb{R}$ with $|\mu| \geqq \sqrt{3}$. In this case we have

$$|d_k^N(\mu)| \geqq |3-\mu|, \qquad k = 0, \cdots, N-1, \tag{4.14.1}$$

$$|d_N^N(\mu)| \geqq 3 \tag{4.14.2}$$

for all N. The first estimate is obviously satisfied for $k=0$. Using $2|\mu| \geqq (9+|\mu|^2)^{1/2} = |3-\mu|$, we obtain from (4.1), assuming that the estimate is already established for k:

$$\begin{aligned} |\operatorname{Im} d_{k+1}^N(\mu)| &= \left| 4 \operatorname{Im}\mu - \frac{9+|\mu|^2}{|d_k^N(\mu)|^2} \operatorname{Im} d_k^N(\mu) \right| \\ &\geqq 4|\mu| - \frac{9+|\mu|^2}{|d_k^N(\mu)|^2} \geqq 4|\mu| - (9+|\mu|^2)^{1/2} \\ &\geqq (9+|\mu|^2)^{1/2}, \qquad k = 0, \cdots, N-2. \end{aligned}$$

This proves (4.14.1). To prove (4.14.2) we note that $\operatorname{Re} d_k^N(\mu)$ is always positive (and decreasing with respect to k) because

$$\operatorname{Re} d_{k+1}^N(\mu) = \frac{9+|\mu|^2}{|d_k^N(\mu)|^2} \operatorname{Re} d_k^N(\mu), \qquad k = 0, \cdots, N-2.$$

Therefore the last equation in (4.1) implies

$$\operatorname{Re} d_N^N(\mu) = 3 + \frac{9+|\mu|^2}{|d_{N-1}^N(\mu)|^2} \operatorname{Re} d_{N-1}^N(\mu) \geqq 3,$$

which proves (4.14.2). Now it follows from (4.14) and (4.7) that

$$|\alpha_N^N(\mu)| = 2\frac{|3-\mu|}{|d_0^N(\mu)|} \cdot \cdots \cdot \frac{|3-\mu|}{|d_{N-1}^N(\mu)|} \cdot \frac{3}{|d_N^N(\mu)|} \leqq 2.$$

It remains to consider $\mu = i\theta$ with $|\theta| < \sqrt{3}$, $\theta \in \mathbb{R}$.

Then we obtain from (4.9), with $k = N$,

$$|\alpha_N^N(\mu)|^2 = 36w^2 \Big/ \left((9+w^2)^2 \frac{1-\cos N\delta}{2} + 36w^2 \frac{1+\cos N\delta}{2} \right) \leqq 1,$$

because $6w \leqq 9 + w^2$. □

From Lemma 4.3 we obtain the following a priori bounds for the unstable eigenvalues of the matrices A^N and $(A_T^N)^*$.

PROPOSITION 4.4. *Let* $\omega = \|A_0\| + 2\|A_1\|$. *For every* $N \in \mathbb{N}$ *and every* $\lambda \in \mathbb{C}$ *with* $\operatorname{Re} \lambda \geqq 0$ *and* $\det \Delta^N(\lambda) = 0$ *we have* $|\lambda| \leqq \omega$.

Proof. It follows from Lemma 4.3 that

$$\left\| A_0 + A_1 \alpha_N^N\left(\frac{\lambda h}{N}\right) \right\| \leqq \|A_0\| + 2\|A_1\| = \omega$$

for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ and every $N \in \mathbb{N}$. Therefore we obtain from (3.10) that $\|\lambda I - \Delta^N(\lambda)\| \leqq \omega$ for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ and every $N \in \mathbb{N}$. Hence $\det \Delta^N(\lambda) \neq 0$ for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ and $|\lambda| > \omega$. This proves the statement of the proposition. □

Now we are in a position to prove the desired result on stabilizability and detectability for the approximating systems (Σ^N).

4.3. Stability, stabilizability, and detectability.

THEOREM 4.5. *The following statements are true*:

(a) *If system* (Σ) *is stable, then there exists an* N_0 *such that system* (Σ^N) *is stable for every* $N \geqq N_0$.

(b) *If system* (Σ) *is stabilizable* (*respectively, detectable*) *then there exists an* N_0 *such that system* (Σ^N) *is stabilizable* (*respectively, detectable*) *for every* $N \geqq N_0$.

Proof. Suppose (Σ) is stable. Then $\det \Delta(\lambda) \neq 0$ for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$. Hence the uniform convergence result for $\Delta^N(\lambda)$ on bounded domains (Theorem 4.2) shows that $\det \Delta^N(\lambda) \neq 0$ for $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ and $|\lambda| \leqq \omega$ provided N is sufficiently large. If $\omega > 0$ is large enough then we obtain from Proposition 4.4 that $\det \Delta^N(\lambda) \neq 0$ for all $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ provided N is sufficiently large. Now the stability of (Σ^N) follows from Theorem 3.8(a). This proves (a). Statement (b) can be established analogously. □

Now we might ask whether the stability of system (Σ) implies stability of the approximating systems (Σ^N) uniformly with respect to N, i.e., the existence of constants $M \geqq 1$, $\varepsilon > 0$ such that

$$\|e^{A^N t}\|_N \leqq M e^{-\varepsilon t}, \qquad t \geqq 0$$

for N sufficiently large. A result of this type would be needed to apply a result of Gibson [8] concerning the approximation of the solution to the algebraic Riccati equation. Moreover, the uniform stability has been stated as a conjecture in [3] for the spline approximation scheme developed in [2]. Our result below shows that such a conjecture is definitely wrong for the approximation scheme developed in this paper. This also indicates that it is wrong for the spline approximation scheme in [2].

PROPOSITION 4.6. *Suppose that there exist constants* $M \geqq 1$ *and* $\varepsilon_N > 0$ *such that*

$$\|\exp(Na^N t)\|_N \leqq M e^{-\varepsilon_N t}, \qquad t \geqq 0,$$

for all N. Then $\varepsilon_N = o(1/N^{1/2})$. *Here* $\|\cdot\|_N$ *denotes the operator norm corresponding to the vector norm* $|x|_N^2 = (1/N) x^T q^N x$ *on* $\mathbb{R}^{N+1}$.

Proof. First note that

$$\tfrac{1}{6}x^T x \leqq x^T q^N x \leqq x^T x, \qquad x \in \mathbb{R}^{N+1},$$

and therefore

$$x^T x \leqq x^T (q^N)^{-1} x \leqq 6 x^T x, \qquad x \in \mathbb{R}^{N+1}.$$

This implies, for $x_0 = \operatorname{col}(1, 0, \cdots, 0)$ and $\mu \in i\mathbb{R}$ (cf. (3.9)),

$$\begin{aligned}
\sum_{k=0}^{N} |\alpha_k^N(\mu)|^2 &\leqq 6N |\alpha^N(\mu)|_N^2 = 6N |(\mu q^N - h^N)^{-1} x_0|_N^2 \\
&= 6N \left| \int_0^\infty e^{-\mu t} \exp(N a^N t/N)(q^N)^{-1} x_0 \, dt \right|_N^2 \\
&= 6NM^2 |(q^N)^{-1} x_0|^2 \left(\int_0^\infty e^{-\varepsilon_N(t/N)} \, dt \right)^2 \\
&= \frac{6N^3 M^2}{\varepsilon_N^2} |(q^N)^{-1} x_0|_N^2 = \frac{6N^2 M^2}{\varepsilon_N^2} x_0^T (q^N)^{-1} x_0 \\
&\leqq \frac{36 N^2 M^2}{\varepsilon_N^2}.
\end{aligned}$$

Therefore

$$\varepsilon_N \leqq 6NM \left(\sum_{k=0}^{N} |\alpha_k^N(\mu)|^2 \right)^{-1/2} \tag{4.15}$$

for all $\mu \in i\mathbb{R}$.

Now let $\mu = i\theta$ satisfy $|\theta| < \sqrt{3}$, $\theta \in \mathbb{R}$. Since $\delta(i\theta) \to 0$ as $|\theta| \to \sqrt{3}$, we can choose a sequence $\theta_N \in \mathbb{R}$, $|\theta_N| < \sqrt{3}$, such that $|\theta_N| \to \sqrt{3}$ and $\delta_N = \delta(i\theta_N) = 2\pi/N$. We put $w_N = w(i\theta_N)$ and get, using (4.9) and $\sum_{k=0}^{N} \cos(2k\pi/N) = 1$,

$$\begin{aligned}
\sum_{k=0}^{N} \left| \alpha_{N-k}^N(i\theta_N) \right|^2 &= \frac{9}{w_N^2} \sum_{k=0}^{N} \left(1 - \cos \frac{2k\pi}{N} \right) + \frac{1}{2} \sum_{k=0}^{N} \left(1 + \cos \frac{2k\pi}{N} \right) \\
&\geqq \frac{9}{2w_N^2} N.
\end{aligned} \tag{4.16}$$

From (4.6) we get

$$\sin \delta(\theta) = \frac{4\theta w(i\theta)}{9 + \theta^2},$$

which shows that for positive constants c_1, c_2,

$$\frac{c_1}{N} \leqq w_N \leqq \frac{c_2}{N} \quad \text{for all } N.$$

This and (4.16) imply

$$\sum_{k=0}^{N} |\alpha_k^N(\mu_N)|^2 \geqq \text{const. } N^3.$$

This last estimate and (4.15) show that

$$\varepsilon_N \leqq \text{const.} \frac{1}{N^{1/2}}.$$

□

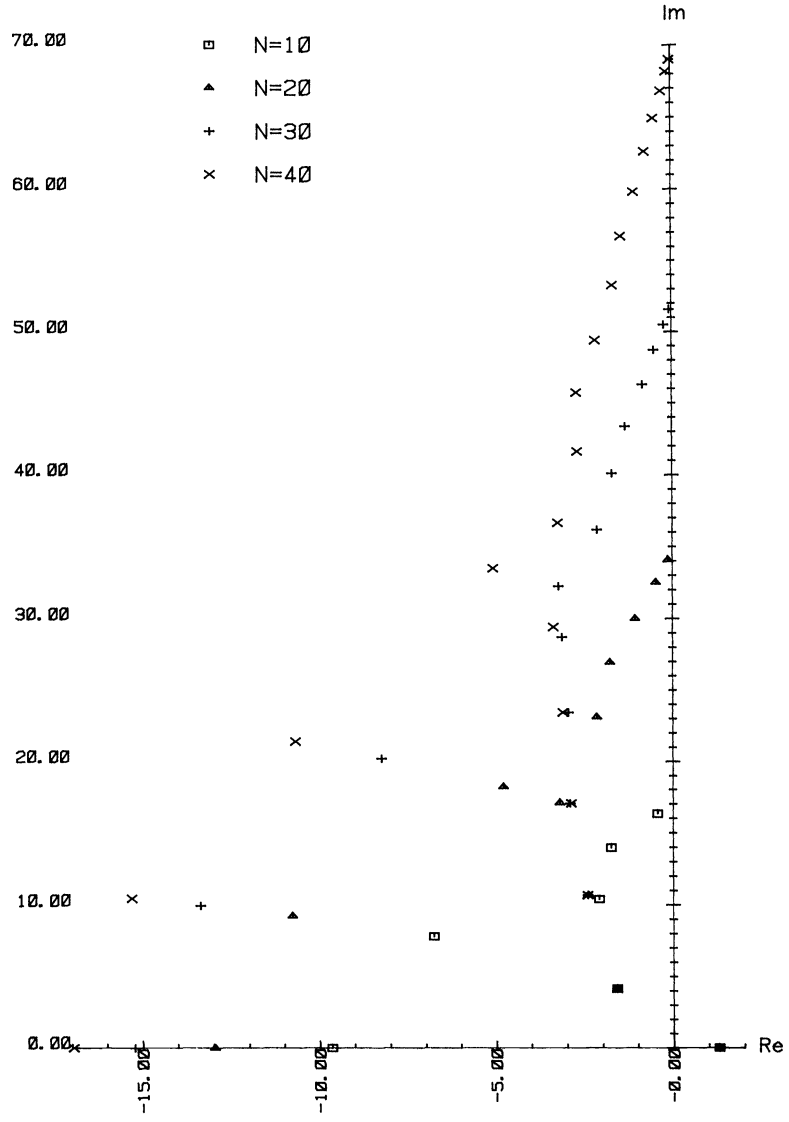
Im
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
N=10
N=20
N=30
N=40
-15.00
-10.00
-5.00
-0.00
Re

FIG. 1

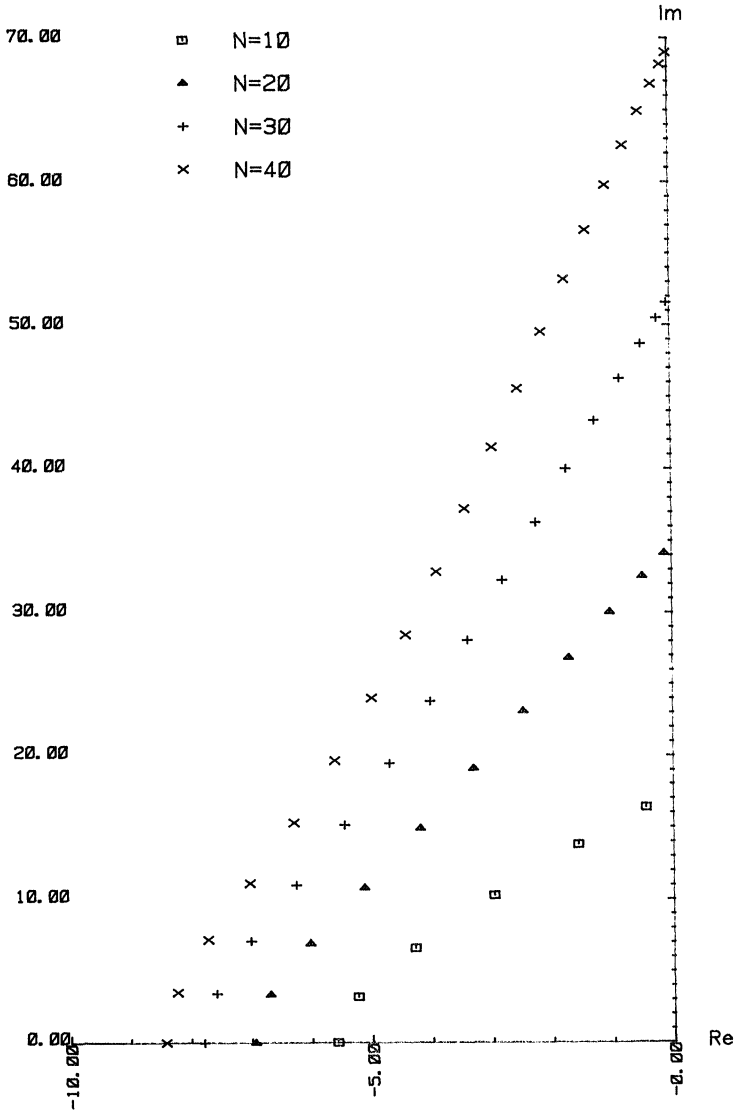
70.00
60.00
50.00
40.00
30.00
20.00
10.00
0.00
N=10
N=20
N=30
N=40
Im
Re
-10.00
-5.00
-0.00

FIG. 2

The result above shows that exponential stability uniform with respect to N is, in general, impossible for our scheme. Numerical studies show that there is a sequence λ_N, $N=1,2,\cdots$, of eigenvalues $\lambda_N \in \sigma((N/h)a^N)$ such that $\operatorname{Re}\lambda_N \to 0$ and $\operatorname{Im}\lambda_N \to \infty$. In fact, the numerical results indicate $\operatorname{Re}\lambda_N = O(1/N^2)$. In the general case where $\sigma((N/h)a^N)$ is not part of $\sigma(A^N)$, numerical studies still show the existence of a sequence λ_N, $N=1,2,\cdots$, such that $\lambda_N \in \sigma(A^N)$ with $\operatorname{Re}\lambda_N < 0$, $\operatorname{Re}\lambda_N \to 0$ and $\operatorname{Im}\lambda_N \to \infty$. Figure 1 illustrates the location of the spectrum for the approximating systems in case of the scalar equation $\dot{x}(t) = x(t) + x(t-1)$. For comparison Fig. 2 illustrates the spectrum of Na^N.

5. Uniform output stability. Despite the negative result of Proposition 4.6 and the fact that some eigenvalues of the approximating systems approach the imaginary axis, we are still able to prove a uniform L^2-estimate for the $\mathbb{R}^n$-components

$$z_0^N(t;\phi) \in \mathbb{R}^n, \qquad w_0^N(t;f) \in \mathbb{R}^n$$

of the unique solutions of (Σ^N) and (Σ_T^{N*}) (with $u \equiv 0$). We call this property the uniform output stability of the systems (Σ^N) and (Σ_T^{N*}).

THEOREM 5.1. *Suppose system* (Σ) *is stable. Then the approximating systems* (Σ^N) *and* (Σ_T^{N*}) *are uniformly output stable for* N *sufficiently large, i.e., there exists an* $N_0 \in \mathbb{N}$ *and a constant* $c > 0$ *such that for* $N \geqq N_0$

$$\int_0^\infty |z_0^N(t;\phi)|_{\mathbb{R}^n}^2 \, dt \leqq c\|\phi\|_{M^2}^2 \quad \text{for all } \phi \in M^2, \tag{5.1}$$

$$\int_0^\infty |w_0^N(t;f)|_{\mathbb{R}^n}^2 \, dt \leqq c\|f\|_{M^2}^2 \quad \text{for all } f \in M^2. \tag{5.2}$$

Proof. Choose $N_0 \in \mathbb{N}$ such that $\det \Delta^N(\lambda) \neq 0$ for all $\lambda \in \mathbb{C}$ with $\operatorname{Re}\lambda \geqq 0$ and all $N \geqq \mathbb{N}_0$ (Theorem 4.5). Then it follows from (3.17), (3.18), and Lemma 3.2 that the Fourier transforms of $z_0^N(t;\phi)$ and $w_0^N(t;f)$ (determined to be identically zero for $t<0$) are given by

$$\hat{z}_0^N(i\omega;\phi) = \Delta^N(i\omega)^{-1} E^N(i\omega)^T Q^N \pi^N F\phi, \qquad \omega \in \mathbb{R},$$

$$\hat{w}_0^N(i\omega;\phi) = \Delta^N(i\omega)^{-1} E^N(i\omega)^T Q^N \pi^N f, \qquad \omega \in \mathbb{R},$$

for $N \geqq N_0$. Using Plancherel's theorem, we see that to prove (5.1) and (5.2) it is enough to show

$$\int_{-\infty}^\infty |\Delta^N(i\omega)^{-1} E^N(i\omega)^T Q^N z|_{\mathbb{R}^n}^2 \, dt \leqq 2\pi c z^T Q^N z$$

for all $z \in \mathbb{R}^{k(N)}$ and all $N \geqq N_0$. The definition of $E^N(\lambda)$ in (3.11), together with Lemma 4.3, show that it suffices to prove a uniform estimate of the form

$$\frac{1}{N}\int_{-\infty}^\infty \|\Delta^N(i\omega)^{-1}\|^2 \sum_{k=0}^N \left|\alpha_k^N\left(\frac{i\omega h}{N}\right)\right|^2 d\omega \leqq c \tag{5.3}$$

for all $N \geqq N_0$ (with a possibly different constant c). Of course, it is only necessary to consider $\omega \geqq 0$.

Using (3.10) and Lemma 4.3 we immediately get the estimate

$$\|\Delta^N(i\omega)^{-1}\| \leqq \frac{1}{|\omega| - c_0} \quad \text{for } |\omega| > c_0, \quad N = 1, 2, \cdots,$$

where $c_0 = \|A_0\| + 2\|A_1\|$. By Theorem 4.2 and the stability assumption on (Σ) we obtain

$$\|\Delta^N(i\omega)^{-1}\|^2 \leqq \frac{c_1}{1+\omega^2} \quad \text{for all } \omega \in \mathbb{R} \tag{5.4}$$

and N sufficiently large, where c_1 is not dependent on N.

Defining

$$f^N(\theta) = \frac{1}{N} \sum_{k=0}^{N} |\alpha_k^N(i\theta)|^2,$$

we find that for all $N = 1, 2, \cdots$

$$f^N(\theta) \leqq \frac{N+1}{N} 4 \leqq 8 \quad \text{for } \theta \geqq \sqrt{3} \tag{5.5}$$

and for any $\alpha \in (0, 1)$

$$f^N(\theta) \leqq \frac{N+1}{N} \frac{2}{1-\alpha^2} \leqq \frac{4}{1-\alpha^2} \quad \text{for } 0 \leqq \theta \leqq \alpha\sqrt{3}. \tag{5.6}$$

The estimate (5.5) is a straightforward consequence of (4.7) and estimates (4.14). To obtain (5.6) we can use the representation (4.9) and the estimates $w^2(i\theta) = 9 - 3\theta^2 \geqq 9(1-\alpha^2)$, $w^2(i\theta) \leqq 9$, and $9 + w^2(i\theta) \geqq 6w(i\theta)$ for $0 \leqq \theta \leqq \alpha\sqrt{3}$.

It remains to investigate the behaviour of $f^N(\theta)$ at intervals of the form $(\alpha\sqrt{3}, \sqrt{3})$, $0 < \alpha < 1$. There we cannot expect to have a bound for $f^N(\theta)$ uniformly with respect to N. Formula (4.9) shows that we should expect difficulties for those θ near $\sqrt{3}$ such that $N\delta(\theta)$ is close to an integer multiple of 2π. This reflects the fact that the eigenvalues of a^N are closest to the imaginary axis near $\pm iN\sqrt{3}$ (see Fig. 2), i.e., for $\theta = \omega h / N$ close to $\pm\sqrt{3}$. In Fig. 3 we show the plot for $f^N(\omega h/N)$, $N = 10, 20, 30, 40$, $h = 1$, which illustrates the difficulties.

We first determine those parts of $(\alpha\sqrt{3}, \sqrt{3})$, where we still can find a uniform bound for $f^N(\theta)$. Since α is not yet fixed we consider $\theta \in [0, \alpha\sqrt{3}]$.

CLAIM 1. *If $\theta \in [0, \sqrt{3}]$ is such that $0 \leqq \delta(\theta) \leqq \pi/3N$, then*

$$f^N(\theta) \leqq 8 \quad \textit{for all } N. \tag{5.7}$$

Proof. From $\frac{1}{2} \leqq \cos N\delta(\theta) \leqq \cos k\delta(\theta)$, $k = 0, \cdots, N$, and (4.9) we get

$$|\alpha_{N-k}^N(i\theta)|^2 \leqq 4 \frac{9(1-\cos N\delta(\theta)) + w^2(i\theta)(1+\cos k\delta(\theta))}{9(1-\cos N\delta(\theta)) + 4w^2(i\theta)(1+\cos N\delta(\theta))}$$

$$\leqq 4 \frac{9(1-\cos N\delta(\theta)) + 2w^2(i\theta)}{9(1-\cos N\delta(\theta)) + 6w^2(i\theta)} \leqq 4,$$

for $k = 0, \cdots, N$, which implies the result. □

CLAIM 2. *If $\theta \in [0, \sqrt{3}]$ is such that $|\delta(\theta) - 2\pi(\nu/N)| \geqq \pi/3N$ for $\nu = 0, \cdots, [N/2]$, then*

$$f^N(\theta) \leqq 64 \quad \textit{for all } N. \tag{5.8}$$

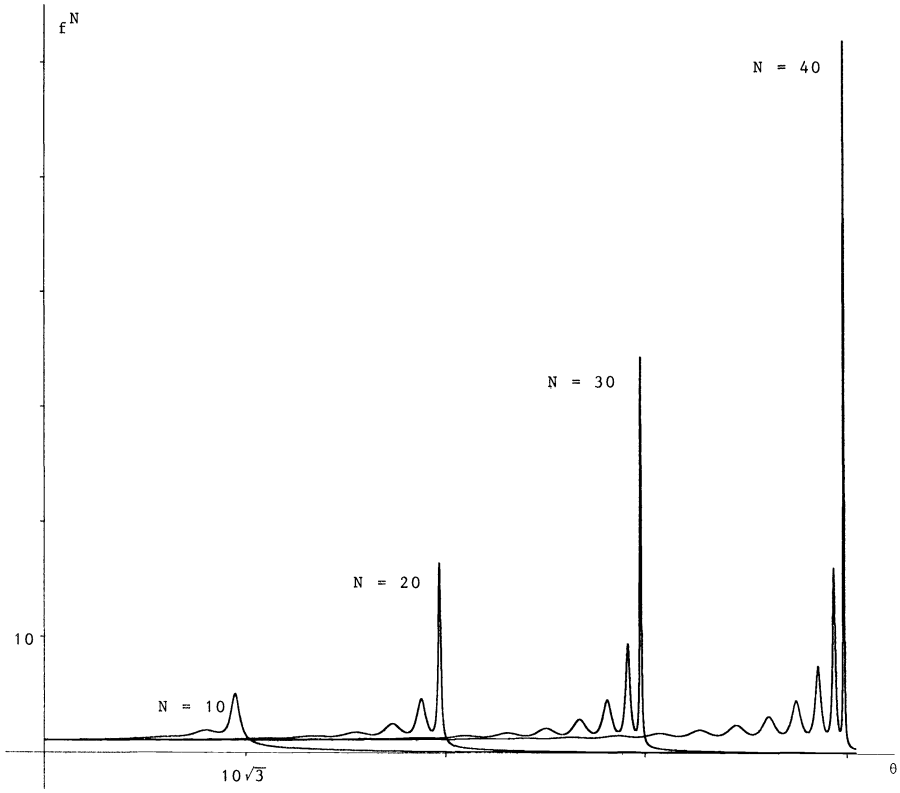

FIG. 3

Proof. From $\cos N\delta(\theta) \leqq \cos(\pi/3) = \frac{1}{2}$, $w^2(i\theta) \leqq 9$ and (4.9) we obtain

$$|\alpha_{N-k}^N(i\theta)|^2 \leqq 36 \frac{2 \cdot 9 + 2 \cdot 9}{9^2(1 - \cos N\delta(\theta))} \leqq 36, \qquad k = 0, \cdots, N. \qquad \square$$

For ν, $N \in \mathbb{N}$ and $\alpha \in (0, 1)$ we define the intervals

$$I_\nu^N = \left\{ \theta \in [\alpha\sqrt{3}, \sqrt{3}] \,\middle|\, |N\delta(\theta) - 2\pi\nu| < \frac{\pi}{3} \right\}.$$

We have $I_\nu^N \neq \varnothing$ if and only if $\nu = 0, \cdots, \nu_1$, where $\nu_1 < N$ is determined by the conditions $2\pi\nu_1 - \pi/3 < N\delta(\alpha\sqrt{3})$ and $2\pi(\nu_1 + 1) - \pi/3 \geqq N\delta(\alpha\sqrt{3})$. Inequalities (5.5)-(5.8) imply that for any $\alpha \in (0, 1)$ there exists a constant $c = c(\alpha)$ independent of N such that

$$f^N(\theta) \leqq c(\alpha), \qquad \theta \in M := [0, \infty) \Big\backslash \bigcup_{\nu=1}^{\nu_1} I_\nu^N \tag{5.9}$$

for all $N \in \mathbb{N}$. Let $\tilde{M} = \{\omega \geqq 0 \,|\, \omega h/N \in M\}$. Then by (5.4) and (5.9)

$$\int_{\tilde{M}} \|\Delta^N(i\omega)^{-1}\|^2 f^N\left(\frac{\omega h}{N}\right) d\omega \leqq c(\alpha) c_1 \frac{\pi}{2}. \tag{5.10}$$

Let $\tilde{I}_\nu^N = \{\omega \geqq 0 \,|\, (\omega h/N) \in I_\nu^N\}$, ν, $N \in \mathbb{N}$. Then by (5.4) (note, that $\omega > \alpha_0\sqrt{3}(N/h)$ for $\omega \in \tilde{I}_\nu^N$)

$$J = \sum_{\nu=1}^{\nu_1} \int_{\tilde{I}_\nu^N} \|\Delta^N(i\omega)^{-1}\|^2 f^N\left(\frac{\omega h}{N}\right) d\omega$$

$$\leqq \frac{c_1}{1+3\alpha_0^2 N^2/h^2} \sum_{\nu=1}^{\nu_1} \frac{N}{h} \int_{I_\nu^N} f^N(\theta)\, d\theta \tag{5.11}$$

for N sufficiently large. It remains to prove an estimate of the form

$$\sum_{\nu=1}^{\nu_1} \frac{N}{h} \int_{I_\nu^N} f^N(\theta)\, d\theta \leqq c$$

for all $N = 1, 2, \cdots$, where c is independent of N.

Now we fix $\alpha = \alpha_0$ by imposing the condition

$$\delta(\alpha_0\sqrt{3}) = \frac{\pi}{2}. \tag{5.12}$$

Then there exists a constant c_2 such that

$$\frac{1}{c_2}\delta(\theta) \leqq w(i\theta) \leqq c_2\delta(\theta), \qquad \theta \in [\alpha_0\sqrt{3}, \sqrt{3}]. \tag{5.13}$$

Proof of (5.13). From (4.6) we see that $\sin \delta(\theta) = 4\theta w(i\theta)/(9+\theta^2)$. Therefore

$$\frac{4\alpha_0\sqrt{3}}{12} w(i\theta) \leqq \sin \delta(\theta) \leqq \frac{4\sqrt{3}}{9} w(i\theta), \qquad \theta \in [\alpha_0\sqrt{3}, \sqrt{3}]. \tag{5.14}$$

The monotonicity of $\delta(\theta)$ and (5.11) imply $0 \leqq \delta(\theta) \leqq \pi/2$. Then $(2/\pi)\delta(\theta) \leqq \sin \delta(\theta) \leqq \delta(\theta)$, which together with (5.14) implies the result. □

CLAIM 3. *There exists a constant c_3 independent of N and ν such that*

$$|I_\nu^N| \leqq c_3 \frac{\nu}{N^2} \tag{5.15}$$

for $N = 1, 2, \cdots$ and $\nu = 1, \cdots, \nu_1$. Here $|I_\nu^N|$ denotes the length of the interval I_ν^N.

Proof. From (5.13) we get

$$-\frac{1}{w(i\theta)} \leqq -\frac{1}{c_2\delta(\theta)}, \qquad \theta \in [\alpha_0\sqrt{3}, \sqrt{3}]. \tag{5.16}$$

The definition (4.6) of $\delta(\theta)$ together with (4.3) implies

$$\delta'(\theta) = -\frac{36}{(9+\theta^2)w(i\theta)}.$$

Using (5.16), for $\theta \in I_\nu^N$ we obtain

$$\delta'(\theta) \leqq -\frac{36}{9+\theta^2} \cdot \frac{1}{c_2\delta(\theta)} \leqq -\frac{3}{c_2\delta(\theta)} \leqq -\frac{1}{c_2\pi} \frac{N}{\nu}.$$

Therefore

$$\frac{2\pi}{3N} = -\int_{I_\nu^N} \delta'(\theta)\, d\theta \geqq \frac{1}{c_2\pi} \frac{N}{\nu} |I_\nu^N|$$

for $N = 1, 2, \cdots$ and $\nu = 1, \cdots, \nu_1$, which proves the result. □

CLAIM 4. *There exists a constant c_4 independent of N and ν such that*

$$f^N(\theta) \leqq c_4 \left(\frac{N}{\nu}\right)^2 \tag{5.17}$$

for $\theta \in I_\nu^N$, $N = 1, 2, \cdots$ and $\nu = 1, \cdots, \nu_1$.

Proof. Let $\theta \in I_\nu^N$. Then $\delta(\theta) > (2\pi\nu - \pi/3)/N$ and by (5.13) we obtain

$$w(i\theta) > \frac{1}{c_2} \cdot \frac{1}{N}\left(2\pi\nu - \frac{\pi}{3}\right) \geqq \frac{5\pi}{3c_2} \cdot \frac{\nu}{N}.$$

This together with (4.9), $w^2(i\theta) \leqq 9$, and $\cos N\delta(\theta) > \frac{1}{2}$ for $\theta \in I_\nu^N$ implies

$$|\alpha_{N-k}^N(i\theta)|^2 \leqq \frac{18}{(1+\cos N\delta(\theta))w^2(i\theta)} \leqq \frac{108c_2^2}{25\pi^2} \cdot \frac{N^2}{\nu^2}.$$

Then the result follows immediately. □

Using (5.11), (5.15), and (5.17) we get the following estimate for J:

$$\begin{aligned} J &\leqq \frac{c_1c_3c_4}{1+3\alpha_0^2N^2/h^2} \frac{N}{h} \sum_{\nu=1}^{\nu_1} \left(\frac{N}{\nu}\right)^2 \frac{\nu}{N^2} \\ &\leqq \frac{c_1c_3c_4}{h} \frac{N}{1+3\alpha_0^2N^2/h^2} \sum_{\nu=1}^{N} \frac{1}{\nu} \leqq \frac{c_1c_2c_3}{h} \frac{N(1+\ln N)}{1+3\alpha_0^2N^2/h^2} \end{aligned}$$

for N sufficiently large. This together with (5.10) establishes (5.3). Thus the proof of Theorem 5.1 is finished.

Remarks. (1) Uniform output stability in general does not imply a uniform (with respect to N) exponential decay for the $\mathbb{R}^n$-components of solutions of the approximating equations. If we are willing to accept the existence of eigenvalues λ_N for the approximating equations with $\operatorname{Re}\lambda_N \to 0$ as $N \to \infty$ also in case $\det A_1 \neq 0$ (as is demonstrated numerically in Fig. 1 but not proved in this paper), it is sufficient to show that in case $\det A_1 \neq 0$ any eigenvector for the approximating system has a nonzero $\mathbb{C}^n$-component. To prove this, assume $y^N = \operatorname{col}(y_0^N, y_1^N)$ with $y_0^N \in \mathbb{C}^n$, $y_1^N \in \mathbb{C}^{(N+1)n}$ is an eigenvector of A^N corresponding to the eigenvalue λ_N. Assume $y_0^N = 0$. Then $A^N y^N = \lambda_N y^N$ is equivalent to

$$(0 \text{ ——— } 0\ A_1)y_1^N = 0 \quad \text{and} \quad \frac{N}{h}(a^N \otimes I)y_1^N = \lambda_N y_1^N. \tag{5.18}$$

The second equation implies $y_1^N = x \otimes v$, where $a^N x = (\lambda h/N)x$, $x = (x_0, \cdots, x_N) \in \mathbb{C}^{N+1}\backslash\{0\}$ and $v \in \mathbb{C}^n\backslash\{0\}$. By Lemma 3.5(b) we have $x_N \neq 0$. The first equation in (5.18) implies $A_1(x_N v) = x_N A_1 v = 0$, a contradiction to $\det A_1 \neq 0$.

(2) It is interesting to state a consequence of uniform output stability for the eigenvectors of the approximating equations. Assume (Σ) is stable so that (5.1) is true, and let $y^N = (y_0^N, y_1^N)$, $y_0^N \in \mathbb{C}^n$, $y_1^N \in \mathbb{C}^{(N+1)n}$ be an eigenvector of A^N corresponding to an eigenvalue λ_N. Then $z_0^N(t; \iota^N y^N) = y_0^N e^{\lambda_N t}$, $t \geqq 0$, and therefore

$$\int_0^\infty |z_0^N(t; \iota^N y^N)|_{\mathbb{C}^n}^2 \, dt = \frac{1}{2|\operatorname{Re}\lambda_N|} |y_0^N|_{\mathbb{C}^n}^2.$$

For $N \geqq N_0$, (5.1) implies

$$|y_0^N|_{\mathbb{C}^n}^2 \leqq \frac{2c|\operatorname{Re}\lambda_N|}{1-2c|\operatorname{Re}\lambda_N|} |(\iota^N y^N)^1|_{L^2}^2$$

provided $|\operatorname{Re}\lambda_N| < 1/2c$. Therefore if $|\operatorname{Re}\lambda_N| \to 0$ as $N \to \infty$ then also $y_0^N \to 0$ in $\mathbb{C}^n$.

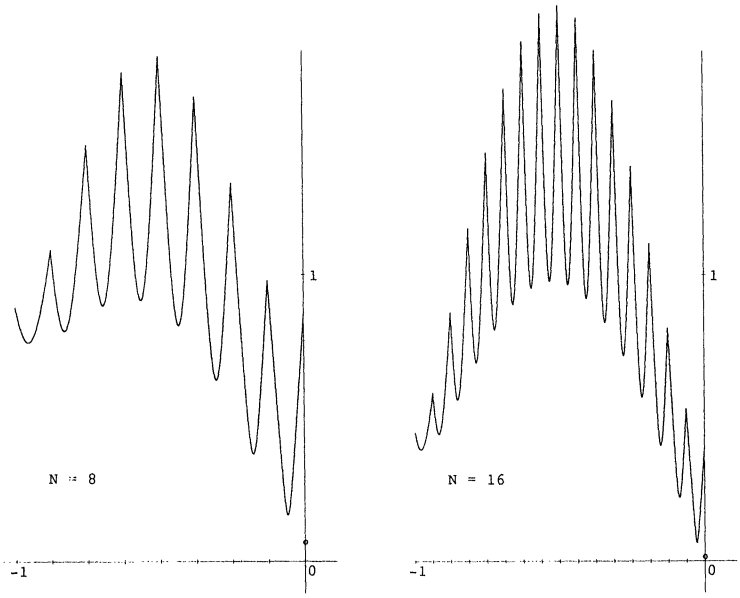

FIG. 4

Figure 4 shows, for $N=8$, 16, $|(\iota^N y^N)^1(\theta)|$, $-1 \leqq \theta \leqq 0$, and $|(\iota^N y^N)^0|$ for the normalized eigenvector y^N (i.e., $\|\iota^N y\|_{M^2}=1$) of A^N corresponding to the eigenvalue λ_N with the smallest real part (and at the same time largest imaginary part) in case of the scalar equation $\dot{x}(t)=-2x(t)+x(t-1)$. The eigenvalues λ_N and $|(\iota^N y^N)^0|$, $N=4$, 8, 16, are given by

$$\lambda_4=-2.9294+5.1788i, \quad \lambda_8=-0.7218+12.7848i, \quad \lambda_{16}=-0.1874+27.0462i,$$
$$|(\iota^4 y^4)^0|=0.3536, \quad |(\iota^8 y^8)^0|=0.0690, \quad |(\iota^{16} y^{16})^0|=0.0167.$$

(3) Uniform output stability is sufficient to prove convergence of the approximating Riccati operators in the case of the infinite time horizon problem observed numerically in [11]. This will be shown in a forthcoming paper [12].

Acknowledgment. We thank W. Prager for the computations concerning Figs. 1-4.

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 AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496-522.

[3] 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.

[4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences 8, Springer-Verlag, Berlin, New York, 1978.

[5] M. C. DELFOUR, E. B. LEE, AND A. MANITIUS, *F-reduction of the operator Riccati equation*, Automatica, 14 (1978), pp. 385-395.

[6a] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems, Part* I; J. Math. Anal. Appl., 73 (1980), pp. 466-490.

[6b] ———, *The structural operator F and its role in the theory of retarded systems, Part* II, J. Math. Anal. Appl., 74 (1980), pp. 359-381.

[7] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537–565.
[8] ———, *Linear quadratic optimal control of hereditary differential systems: Infinite dimensional Riccati equation and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95–139.
[9] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, New York, 1977.
[10] F. KAPPEL AND D. SALAMON, *On the stability properties of spline approximations for retarded systems*, Technical Report No. 78–1986, Institute for Mathematics, University of Graz, Graz, Austria, 1986.
[11] ———, *Spline approximation for retarded systems and the Riccati equation*, SIAM J. Control Optim., 25 (1987), pp. 1082–1117.
[12] ———, *An approximation theorem for the algebraic Riccati equation*, SIAM J. Control Optim., submitted.
[13] A. MANITIUS, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.
[14] ———, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, SIAM J. Control Optim., 19 (1981), pp. 516–532.
[15] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, Berlin, New York, 1983.
[16] D. SALAMON, *On dynamic observation and state feedback for time delay systems*, in Evolution Equations and their Applications, F. Kappel and W. Schappacher, eds., Research Notes in Mathematics 68, Pitman, London, 1982, pp. 202–219.
[17] ———, *On controlllability and observability of time delay systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 432–439.
[18] ———, *Control and Observation of Neutral Systems*, Research Notes in Mathematics 91, Pitman, London, 1984.
[19] ———, *Structure and stability of finite dimensional approximations for functional differential equations*, SIAM J. Control Optim., 23 (1985), pp. 928–951.
[20] 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. 27, No. 2, pp. 432-445, March 1989

010

MINIMIZING ESCAPE PROBABILITIES: A LARGE DEVIATIONS APPROACH*

PAUL DUPUIS† AND HAROLD KUSHNER‡

Abstract. This paper considers the problem of controlling a possibly degenerate diffusion process so as to minimize the probability of escape over a given time interval. It is assumed that the control acts on the process through the drift coefficient, and that the noise coefficient is small. Developing a large deviations type of theory for the controlled diffusion produces several results. The limit of the normalized log of the minimum exit probability is identified as the value I of an associated (deterministic) differential game. Furthermore, we identify a deterministic (and ε-independent) mapping g from the sample values $\varepsilon w(s)$, $0 \leqq s \leqq t$, into the control space such that if we define the control used at time t by $u(t) = g(\varepsilon w(s), 0 \leqq s \leqq t)$, then the resulting control process is progressively measurable and δ-optimal (in the sense that the limit of the normalized log of the exit probability is within δ of I).

Key words. controlled diffusions, large deviations, differential games

AMS(MOS) subject classifications. primary 93E20, 60F10; secondary 92D25

1. Introduction.

Consider the white-noise-driven control system living in $\mathbb{R}^d$:

$$dx^{u,\varepsilon} = b(x^{u,\varepsilon}, u)\,dt + \varepsilon\sigma(x^{u,\varepsilon})\,dw, \tag{1.1}$$

where u takes values in a compact set $K \subset \mathbb{R}^n$. There are many problems where we want to keep $x^{u,\varepsilon}(\cdot)$ in a set G until some particular job is finished. For example, in the problem of pointing a telescope on a satellite, the domain G and the duration are determined by the object to be photographed and the time required. See Meerkov and Runolfsson [6] for additional examples.

The associated control problem can be formulated in several different ways, depending on the time interval of interest. We consider two criteria. Define $\tau^{u,\varepsilon} = \inf\{t: x^{u,\varepsilon}(t) \in \partial G\}$. One criterion is to minimize

$$P_x\{\tau^{u,\varepsilon} \leqq T\}, \qquad x \in G^0 = \text{interior of } G \tag{1.2a}$$

for given T. The other criterion of interest here is the maximization of

$$E_x\tau^{u,\varepsilon}, \qquad x \in G^0. \tag{1.2b}$$

P_x and E_x denote the probability and expectation (respectively) given $x^{u,\varepsilon}(0) = x$.

In general, it is very difficult to solve for the optimal control. However, in many problems the parameter ε is small. The theory of large deviations provides an alternative that can give a nearly optimal control for small ε, and a great deal more information and insight into the control process, likely escape routes, error bounds, etc. Take u to be a feedback function $u(x, t)$ that is smooth in x, uniformly in $t \leqq T$. Let r be the

* Received by the editors December 23, 1987; accepted for publication (in revised form) April 11, 1988.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003. This author's research was supported in part by National Science Foundation grant DMS-8511470 and in part by Army Research Office grant DAAL03-86-K-0171.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This author's research was supported in part by National Science Foundation grant ECS-8505674, in part by Air Force Office of Scientific Research grant AFOSR-85-0315, and in part by Army Research Office grant DAAL03-86-K-0171.

dimension of $w(\cdot)$. For $v(\cdot)$ that is measurable, lives in $\mathbb{R}^r$, and is such that $\int_0^1 |v|\, dt < \infty$, define the system

$$\dot{\phi} = b(\phi, u(\phi, t)) + \sigma(\phi)v, \qquad \phi(0) = x,$$

and define

$$S(x, u, T) = \inf\left\{\frac{1}{2}\int_0^T |v(t)|^2\, dt\colon \phi(t) \in \partial G \text{ for some } t \leqq T\right\}.$$

The theory of large deviations tells us (under some other regularity conditions) that

$$S(x, u, T) = -\lim_{\varepsilon} \varepsilon^2 \log P_x\{\tau^{u,\varepsilon} \leqq T\}.$$

Because of this result, we are tempted to try to maximize (or nearly maximize) $S(x, u, T)$, and to use the corresponding (if any) maximizing (or "smooth" nearly maximizing) control. This approach encounters serious unresolved technical difficulties. In particular, it is not at all clear that the supremum over smooth feedback controls will be as large as that obtained over alternative classes of controls, such as those used below. Note that since we wish to supremize (over u) an infimum over v, the basic problem can be formulated as a differential game.

We mention here that calculating the limit of the normalized log of the minimum exit probability is by itself not useful in establishing the optimal performance for all small ε of any given control scheme. It may happen that a control that is found to be good for a small but fixed $\varepsilon > 0$ actually behaves poorly in the limit $\varepsilon \to 0$. Obtaining a "good" control that depends on ε only through the actual driving noise process will be an important part of the development below.

Known results in this area are few in number. Fleming and Souganidis [3] consider the large deviations problem associated with the minimization of (1.2a) over the class of feedback controls taking values in K. By use of PDE-viscosity solution techniques they calculate the asymptotics of the infimum of the exit probabilities. Their approach is restricted to the case where the diffusion is uniformly nondegenerate: $\sigma(x)\sigma'(x) \geqq cI$, with $c > 0$. Furthermore, they identify the limit as the value of a certain associated (deterministic) differential game. They do not deal with the uniformity issue raised previously, nor with the problem of construction of δ-optimal policies and their uniformity properties. Wentzell and Freidlin [5] consider the optimization problem associated with (1.2b) for a wide class of processes that includes (1.1) in the uniformly nondegenerate case. However, to obtain a solution with the desired properties, they restrict the class of available controls in a way that is probably not natural for these types of problems. For example, they consider feedback controls that are continuous, except possibly at one point. Simple examples in dimension greater than one show that the "best" control may have discontinuities along manifolds of dimension one less.

The objective of this paper is to extend the conclusions of [3]. By use of probabilistic arguments (as opposed to PDE), we recover the results presented there. The probabilistic arguments allow us to extend these results to the important *degenerate case*, which is in fact more natural in applications. We also address the uniformity issue raised above. The results in this direction are not completely satisfactory in that the exhibited control is not of the simple feedback form, but depends on the "full information" of the past. However, they do suggest that feedback controls are available that do not depend on ε explicitly, and are nearly optimal for small ε.

Our basic assumptions and definitions are as follows.

Assumption A1.

(1) $b(\cdot,\cdot)$ and $\sigma(\cdot)$ are Lipschitz with constant $\bar{K}$ and bounded with constant B on an open set containing $\bar{G}$, the closure of G.

(2) The control space K is compact and independent of time.

(3) G is an open set in $\mathbb{R}^d$.

(4) Either (i) $\sigma(\cdot)$ is a square matrix and uniformly nondegenerate, or (ii) we can partition b and σ in the form

$$b(x,u)=\begin{bmatrix} b_1(x,u) \\ b_2(x) \end{bmatrix}, \qquad \sigma(x)=\begin{bmatrix} \sigma_1(x) \\ 0 \end{bmatrix},$$

where $\sigma_1(\cdot)$ is a square matrix and uniformly nondegenerate.

Throughout the paper we shall assume that we are given a probability space $(\Omega, \mathcal{F}, \mathcal{F}(t), P)$ and a Wiener process $w(\cdot)$ on $[0,1]$ with respect to $\mathcal{F}(t)$. We then take as our class of admissible controls the set of K-valued progressively measurable processes. We denote the set of all such processes by F. For convenience we recall the definition of a progressively measurable process (with respect to $\mathcal{F}(t)$).

DEFINITION. A stochastic process $\xi(t)$ on the sample space Ω and time interval $[0,1]$ is $\mathcal{F}(t)$-progressively measurable if the mapping $[0,t]\times\Omega \ni (s,\omega)\to\xi(s)(\omega)$ is $B(t)\times\mathcal{F}(t)$ measurable for every $0\leqq t\leqq 1$, where $B(t)$ is the Borel σ-algebra of $[0,t]$.

Remark. The symbol u will be used to represent two different types of control processes, depending on the context. At times it will be a deterministic process used in the differential game, and at times it will denote a stochastic process used to control the diffusion. Likewise v will be used to represent both stochastic and deterministic processes, depending on the context. In all cases the intended use should be clear.

The organization of the remainder of the paper is as follows. In § 2 we give a precise definition of the associated differential game in terms of an adaptation of the Elliott-Kalton [4] formulation, and discuss how the existence of value for this differential game relates to our problem. The only difference between our definition and the usual Elliott-Kalton definition is the added requirement that the maps α and β defined below must be measurable. The additional requirement of measurability is due to the fact that several uses are made of stochastic processes defined by composing α (or β) with a given progressively measurable process. Measurability of α (or β) ensures that the resulting process is adapted. The addition of this condition does not change the resulting "value" of the game. Section 3 contains the statement and proof of the main theorem. The proofs of several technical lemmas make up a concluding Appendix. For notational simplicity, we shall consider the problem on the interval $[0,1]$. The results carry over to an arbitrary interval in the obvious way.

Notation. We use $C_x[0,1]$ to denote the set of continuous functions taking values in $\mathbb{R}^k$ (with k depending on the context) and starting at x, and take $d(\cdot,\cdot)$ as the sup norm metric in this space.

2. The associated differential game. Define

$$M=\{u:[0,1]\to K: u \text{ is measurable}\},$$

$$N=\left\{v:[0,1]\to\mathbb{R}^r: \int_0^1 |v|\,dt<\infty\right\}.$$

We identify any two functions that agree almost everywhere and consider M and N as metric spaces with the L^1 metric. A mapping $\alpha: N\to M$ is called a *strategy for the maximizing player* if α is measurable (with respect to the Borel σ-algebras induced

by the inherited metric), and if whenever $0 \leqq s \leqq 1$ and

$$v(t) = \hat{v}(t) \quad \text{for a.e. } 0 \leqq t \leqq s,$$

then

$$\alpha[v](t) = \alpha[\hat{v}](t) \quad \text{for a.e. } 0 \leqq t \leqq s.$$

A *strategy for the minimizing player* is defined in an analogous way, and such a strategy will be denoted by the symbol β. The set of all minimizing (respectively, maximizing) strategies will be denoted by Δ (respectively, Γ).

Next define $\chi(x)$ to be zero if $x \in \partial G$ and $+\infty$ if $x \in G^0$. The definition of the differential game (DG) is then given in terms of the following dynamical equation and cost.

Dynamics.

$$\dot{\phi} = b(\phi, u) + \sigma(\phi)v, \qquad \phi(0) = x. \tag{2.1}$$

Let $\tau_x = \inf\{t: \phi(t) \in \partial G\} \wedge 1$.

Cost. For $\phi(\cdot)$ defined through (2.1), set

$$C(u, v) = \frac{1}{2}\int_0^{\tau_x} |v(t)|^2 \, dt + \chi(\phi(\tau_x)).$$

We then define the *lower value* of the DG by

$$I^-(x) = \inf_{\beta \in \Delta} \sup_{u \in M} C(u, \beta[u]).$$

The *upper value* is defined by

$$I^+(x) = \sup_{\alpha \in \Gamma} \inf_{v \in N} C(\alpha[v], v).$$

Remarks. The terms "upper" and "lower" refer to which player has the "information advantage." In a heuristic sense, for the game corresponding to the lower value we allow the minimizing player (v here) to know the next move of the maximizing player (u) before choosing his own move. Although this distinction is somewhat obscured in the abstract Elliott-Kalton formulation, it is intuitively obvious in the Fleming and Friedman formulations [1], which are equivalent to the Elliott-Kalton formulation under some hypotheses. The reader is referred to [1] for further discussion. The DG we consider differs from that of [3], but it seems to be more natural for this type of problem. The remarks that follow illustrate this point.

The Elliott-Kalton definitions of upper and lower values in terms of strategies have interesting interpretations in terms of the large deviations properties of the controlled diffusion. First note that the v-control in the DG plays the role of the small noise $\varepsilon \dot{w}$ in the diffusion. Let small $\delta > 0$ be given. Consider the upper value $I^+(x)$, and let α be a "nearly" optimizing strategy for the maximizing player. Let $v \in N$ be given. Then the "nearly" supremizing α gives us a strategy that accomplishes one of two things. Either $\chi(\phi(\tau_x)) = \infty$ (ϕ never escapes from G) or

$$\frac{1}{2}\int_0^{\tau_x} |v(t)|^2 \, dt \geqq I^+(x) - \delta$$

(ϕ escapes from G, but at a "cost" of not less than $I^+(x) - \delta$). Large deviations theory for the process $\varepsilon \dot{w}$ then suggests that, when ε is small, the probability of $\varepsilon \dot{w}$ "tracking" one of the v functions corresponding to escape from G (in the sense that $x^{u,\varepsilon}$ is near to the corresponding ϕ associated with $\alpha[v], v$) is no greater than $\exp -(I^+(x) - 2\delta)/\varepsilon^2$.

This suggests that we can obtain a progressively measurable control u_0 from the "nearly" supremizing α so that when ε is small,

$$P_x\{\tau^{u_0,\varepsilon} \leqq 1\} \leqq \exp -(I^+(x)-2\delta)/\varepsilon^2.$$

On the other hand, consider the lower value $I^-(x)$, and let β be "nearly" infimizing. Then, no matter what progressively measurable control strategy $u(t)$ is used, β describes a path for the noise to follow whose "action" or "cost" is no greater than $I^-(x)+\delta$, and that leads to escape. The large deviations properties of $\varepsilon\dot{w}$ now suggest that no matter what control is used, the probability of escape should (roughly) be bounded below by $\exp -(I^-(x)-2\delta)/\varepsilon^2$.

Thus we have (roughly)

$$\exp -(I^-(x)-2\delta)/\varepsilon^2 \leqq P_x\{\tau^{u_0,\varepsilon} \leqq 1\} \leqq \exp -(I^+(x)-2\delta)/\varepsilon^2,$$

with the conclusion that $I^-(x) \geqq I^+(x)$. From the definition of the game it is possible to show $I^-(x) \leqq I^+(x)$, which implies that the game has a value.

3. The main theorem. Before stating the main theorem, we introduce a "continuity" assumption on the domain G. Define G^δ for small δ as follows: if $\delta \geqq 0$, then

$$G^\delta = \{x \in \mathbb{R}^d : \inf\{|x-y|: y \in G\} \leqq \delta\};$$

if $\delta < 0$, then

$$G^\delta = \{x \in \mathbb{R}^d : \inf\{|x-y|: y \notin G\} \geqq -\delta\}.$$

Next define $I^+(x, \delta)$, $I^-(x, \delta)$ as the upper and lower values of the DG defined in § 2, but with G^δ replacing G there. Since $I^+(x, \delta)$ (respectively, $I^-(x, \delta)$) is monotone nondecreasing in δ, the set of discontinuities of $I^+(x, \cdot)$ (respectively, $I^-(x, \cdot)$) is countable. (Note that x is fixed here.)

Assumption A2. $I^+(x, \delta)$ and $I^-(x, \delta)$ are continuous at $\delta = 0$.

Remarks. It is simple to prove in the uniformly nondegenerate case that $I^+(x, \cdot)$ and $I^-(x, \cdot)$ are in fact continuous functions. This follows from the fact that $b(\cdot, \cdot)$ is bounded on $G \times K$, while v is allowed to "push" the state in any direction. In the degenerate case it can happen that $I^+(x, \cdot)$ (or $I^-(x, \cdot)$) is in fact discontinuous at $\delta = 0$, but even then Assumption A2 is not very restrictive, since it is satisfied for an arbitrarily small perturbation of G. A consequence of the theorem stated below is that at points at which both $I^+(x, \cdot)$ and $I^-(x, \cdot)$ are continuous, we have $I^+(x, \delta) = I^-(x, \delta)$. Monotonicity then implies that $I^+(x, \cdot)$ and $I^-(x, \cdot)$ have the same set of discontinuity points. It should also be noted that in order to obtain the result analogous to the main theorem in the simpler case of uncontrolled diffusion processes:

$$\lim_\varepsilon \varepsilon^2 \log P_x\{\tau^\varepsilon \leqq 1\} = -I(x),$$

the assumption obtained from Assumption A2 when the set K contains only one element is also required.

THEOREM. *Assume* A1 *and* A2, *and let* $I^+(x)$ *and* $I^-(x)$ *be the upper and lower values of the* DG *described in* § 2. *For any* $u \in F$, *let* $x^{u,\varepsilon}(\cdot)$ *be the solution of*

$$dx^{u,\varepsilon} = b(x^{u,\varepsilon}, u)\,dt + \varepsilon\sigma(x^{u,\varepsilon})\,dw, \qquad x^{u,\varepsilon}(0) = x, \tag{3.1}$$

and define

$$\tau^{u,\varepsilon} = \inf\{t: x^{u,\varepsilon}(t) \in \partial G\}. \tag{3.2}$$

Then

$$\underline{\lim_\varepsilon}\, \varepsilon^2 \log \inf_{u \in F} P_x\{\tau^{u,\varepsilon} \leqq 1\} \geqq -I^-(x), \tag{3.3 (1)}$$

(2) *Given $c>0$ there exists a measurable function $g: C_0[0,1] \to M$ with the following properties*:

(i) *If $0 \leqq s \leqq 1$ and $f(t) = \hat{f}(t)$ for $0 \leqq t \leqq s$, then $g[f](t) = g[\hat{f}](t)$ for $0 \leqq t \leqq s$, almost everywhere*;

(ii) *If we define $u = g[\varepsilon w]$, then $u \in F$ and*

$$\overline{\lim_{\varepsilon}}\, \varepsilon^2 \log P_x\{\tau^{u,\varepsilon} \leqq 1\} \leqq -I^+(x) + c, \tag{3.4}$$

$$I^+(x) = I^-(x). \tag{3}$$

Remarks. Part (2) of the theorem gives the existence of a c-optimal (in the asymptotic sense) control u that depends on $x^{u,\varepsilon}(s)$, $0 \leqq s \leqq t$, at time t. Part (1) yields an important uniformity property. For any given $c>0$ and any (possibly ε-dependent) progressively measurable control u_ε, there is $\varepsilon_0 > 0$ such that for $0 < \varepsilon \leqq \varepsilon_0$,

$$P_x\{\tau^{u_\varepsilon,\varepsilon} \leqq 1\} \geqq P_x\{\tau^{u,\varepsilon} \leqq 1\} \exp - c/\varepsilon^2.$$

Proof of (1). For $c>0$ there exists $\delta > 0$ such that $I^-(x,\delta) \leqq I^-(x) + c$. Consider now the DG with domain G^δ and let $C^\delta(u,v)$ denote the cost associated with the domain G^δ. Then there exists a minimizing strategy $\beta \in \Delta$ such that

$$\sup_{u \in M} C^\delta(u, \beta[u]) \leqq I^-(x) + 2c. \tag{3.5}$$

If we redefine $\beta[u](t)$ to be zero when $t \geqq \tau_x$ (given by (2.1)), then β is still a strategy and obviously still satisfies (3.5).

Without loss of generality we may assume the following property of the chosen strategy $\beta: (d/dt)\beta[u](t)$ exists for all $u \in M$ (almost surely in t) and furthermore there is $C_1 < \infty$ such that

$$\left|\frac{d}{dt}\beta[u](t)\right| \vee |\beta[u](t)| \leqq C_1$$

(almost surely in t) for all $u \in M$. This fact follows from Assumption A2 and Lemma A1 of the Appendix.

Take any control process $u \in F$, and define the processes

$$v(t) = \beta[u](t),$$

$$\dot{\phi}^\varepsilon = b(x^{u,\varepsilon}, u) + \sigma(\phi^\varepsilon)v, \qquad \phi^\varepsilon(0) = x.$$

We then have

$$|\dot{v}(t)| \vee |v(t)| \leqq C_1 \quad \text{(a.s. in } t\text{)}$$

for every ω. It follows from the definition of a strategy that $v(t)$ is $\mathcal{F}(t)$ measurable. Since the β under consideration has the property that $\beta[u](\cdot)$ is continuous for every $u \in M$, $v(\cdot)$ and $\phi^\varepsilon(\cdot)$ are $\mathcal{F}(t)$-progressively measurable processes [7, Thm. 1.5.1].

Now define $y^\varepsilon = x^{u,\varepsilon} - \phi^\varepsilon$. Then y^ε satisfies the stochastic equation

$$dy^\varepsilon = \sigma(x^{u,\varepsilon})\varepsilon\, dw - \sigma(\phi^\varepsilon)v\, dt, \qquad y^\varepsilon(0) = 0. \tag{3.6a}$$

Let P_1 denote the measure induced on $C_0[0,1]$ by the solution to (3.6a). By Girsanov's theorem there is a Brownian motion $\bar{w}(\cdot)$ (with respect to the same filtration $\mathcal{F}(\cdot)$ as $w(\cdot)$) such that

$$dy^\varepsilon = \sigma(x^{u,\varepsilon})\varepsilon\, d\bar{w}, \qquad y^\varepsilon(0) = 0, \tag{3.6b}$$

and such that if P_0 is the measure induced on $C_0[0,1]$ by (3.6b), then

$$\frac{dP_1}{dP_0} = \exp\left[\frac{1}{\varepsilon^2}\int_0^1 \langle \sigma(\phi^\varepsilon)v, \sigma(x^{u,\varepsilon})\varepsilon\, d\bar{w}\rangle - \frac{1}{2\varepsilon^2}\int_0^1 |\sigma^{-1}(x^{u,\varepsilon})\sigma(\phi^\varepsilon)v|^2\, dt\right].$$

(In the degenerate case replace σ by σ_1 in the above.)

Define $\Omega^\varepsilon_{\delta_2} = \{\omega: \sup_{0\le t\le 1}|y^\varepsilon(t)| \le \delta_2\}$. We will use the equality

$$P_1(\Omega^\varepsilon_{\delta_2}) = \int_{\Omega^\varepsilon_{\delta_2}} \frac{dP_1}{dP_0}\, dP_0.$$

First note that for any $\delta_2 > 0$, $P_0(\Omega^\varepsilon_{\delta_2}) \to 1$, as $\varepsilon \to 0$. Using the nondegeneracy and the Lipschitz continuity of $\sigma(\cdot)$ (or of $\sigma_1(\cdot)$ in the degenerate case), for given $\delta' > 0$ there is $\delta'' > 0$ such that $|x-y| \le \delta''$ implies $|\sigma^{-1}(x)\sigma(y) - I| \le \delta'$. This, together with (3.5), yields

$$\frac{1}{2}\int_0^1 |\sigma^{-1}(x^{u,\varepsilon})\sigma(\phi^\varepsilon)v|^2\, dt \le I^-(x) + 3c \tag{3.7}$$

on $\Omega^\varepsilon_{\delta_2}$, if δ_2 is small enough.

Finally, we consider the term

$$\left|\int_0^1 \langle \sigma(\phi^\varepsilon)v, dy^\varepsilon\rangle\right|.$$

Since $(d/dt)\sigma(\phi^\varepsilon(t))v(t)$ is bounded, an integration by parts yields the bound $\delta_2 C_2$ for some fixed finite constant C_2, on the set $\Omega^\varepsilon_{\delta_2}$.

Assembling these estimates, we have (for small enough δ_2)

$$P_1(\Omega^\varepsilon_{\delta_2}) \ge \exp -(I^-(x) + 5c)/\varepsilon^2 \tag{3.8}$$

when ε is small. We now pick δ_2 small enough so that the event $\sup_{0\le t\le 1}|y^\varepsilon(t)| \le \delta_2$ implies $x^{u,\varepsilon}(t)$ exits G before $t=1$. The Lipschitz condition on $b(\cdot,\cdot)$ implies that on $\Omega^\varepsilon_{\delta_2}$,

$$\dot{\phi}^\varepsilon = b(\phi^\varepsilon, u) + \gamma + \sigma(\phi^\varepsilon)v, \qquad \phi^\varepsilon(0) = x,$$

where $\sup_{0\le t\le 1}|\gamma(t)| \le \bar{K}\delta_2$. We compare ϕ^ε to the solution of

$$\dot{\psi} = b(\psi, u) + \sigma(\psi)v, \qquad \psi(0) = x.$$

By Gronwall's lemma, and the various Lipschitz and boundedness conditions, we can pick $\delta_2 \le \delta/2$ so that $d(\phi^\varepsilon, \psi) \le \delta/2$ on $\Omega^\varepsilon_{\delta_2}$. By the definition of β, $\psi(\cdot)$ must exit G^δ before time $t=1$. Hence on $\Omega^\varepsilon_{\delta_2}$ it must happen that $x^{u,\varepsilon}(\cdot)$ exits G before $t=1$. This, combined with (3.8), finishes the proof. □

Proof of (2). Now consider the upper value of the differential game:

$$I^+(x) = \sup_{\alpha\in\Gamma}\inf_{v\in N} C(\alpha[v], v).$$

Fix $c > 0$, and pick $\delta > 0$ so that $I^+(x, -\delta) \ge I^+(x) - c$. Let α be a "nearly" maximizing strategy for the differential game with domain $G^{-\delta}$, in the sense that

$$\inf_{v\in N} C^{-\delta}(\alpha[v], v) \ge I^+(x, -\delta) - c. \tag{3.9}$$

We next describe how we use α to control the diffusion process. Let the Wiener process $w(\cdot)$ be given, and define (for $\Delta > 0$)

$$v^{\Delta}(t) = \begin{cases} 0 & \text{for } t \in [0, \Delta), \\ [w(n\Delta) - w(n\Delta - \Delta)]/\Delta & \text{for } t \in [n\Delta, n\Delta + \Delta), \quad n \geqq 1. \end{cases} \tag{3.10}$$

We then define our control process by

$$u(t) = \alpha[\varepsilon v^{\Delta}](t). \tag{3.11}$$

From Assumption A2 and Lemma A2 of the Appendix it follows that we may assume without loss of generality that the strategy α has been chosen so that $\alpha[v](\cdot)$ is a piecewise constant function for every $v \in N$. As was the case previously, the definition of a strategy implies $u(t)$ is $\mathscr{F}(t)$ measurable. Hence $u(t)$ is an $\mathscr{F}(t)$-progressively measurable process [7, Thm. 1.5.1].

The controlled diffusion is therefore

$$dx^{u,\varepsilon} = b(x^{u,\varepsilon}, u)\, dt + \varepsilon\sigma(x^{u,\varepsilon})\, dw, \qquad x^{u,\varepsilon}(0) = x. \tag{3.12}$$

To prove the desired result it is convenient to compare $x^{u,\varepsilon}(\cdot)$ with the solution to

$$\dot{x}^{\varepsilon,\Delta} = b(x^{\varepsilon,\Delta}, u) + \varepsilon\sigma(x^{\varepsilon,\Delta})v^{\Delta}, \qquad x^{\varepsilon,\Delta}(0) = x.$$

Assume that for any given $\rho > 0$ and $M < \infty$ we can show the existence of $\varepsilon_0 > 0$ and $\Delta_0 > 0$ so that for $\Delta \leqq \Delta_0$, $\varepsilon \leqq \varepsilon_0$

$$P_x\{d(x^{u,\varepsilon}, x^{\varepsilon,\Delta}) \geqq \rho\} \leqq \exp - M/\varepsilon^2. \tag{3.13}$$

Then by taking $M = I^{+}(x) + 1$ and $\rho = \delta$, it is obvious that the upper bound is proved if we can show

$$\overline{\lim}\ \varepsilon^2 \log P_x\{x^{\varepsilon,\Delta}(t) \in \partial G^{-\delta} \text{ for some } t < 1\} \leqq -I^{+}(x, -\delta) + 2c. \tag{3.14}$$

However, this follows from our choice of α. Since (3.9) holds, there are only two possibilities for each $v \in N$. Either

$$\frac{1}{2}\int_0^1 |v(t)|^2\, dt \geqq I^{+}(x, -\delta) - c, \tag{3.15}$$

or the solution of (2.1) *does not* escape $G^{-\delta}$ by time $t = 1$. Hence $x^{\varepsilon,\Delta}(\cdot)$ escapes only on the set of paths for which

$$\frac{\varepsilon\Delta}{2} \sum_{0}^{1/\Delta - 1} v^{\Delta}(i\Delta)^2 = \varepsilon \sum_{1}^{1/\Delta - 1} [w(i\Delta) - w(i\Delta - \Delta)]^2/2\Delta \geqq I^{+}(x, -\delta) - c. \tag{3.16}$$

Standard estimates from the theory of large deviations [2] imply that there exist $\Delta_0 > 0$, $\varepsilon_0 > 0$ such that for $\Delta \leqq \Delta_0$, $\varepsilon \leqq \varepsilon_0$ the probability of the event given in (3.16) is less than $\exp - (I^{+}(x, -\delta) + 2c)/\varepsilon^2$. We are therefore finished, except for the proof of (3.13). The details of this estimate are given in Lemma A3 of the Appendix.

Proof of (3). It follows from (1) and (2) that $I^{-}(x) \geqq I^{+}(x)$. We give the easy proof of $I^{-}(x) \leqq I^{+}(x)$ in Lemma A4 of the Appendix, which completes the proof. □

Appendix. In this Appendix we prove several technical lemmas that are needed to prove the main theorem of § 3. Before presenting the lemmas we introduce some new notation. For $-1 \leqq s \leqq 1$, we define $\Delta(s)$ as the set of all measurable mappings β from $M \to N$ such that

$$u(r) = \hat{u}(r) \quad \text{for a.e. } 0 \leqq r \leqq t$$

implies

$$\beta[u](r) = \beta[\hat{u}](r) \quad \text{for a.e. } 0 \leqq r \leqq \min(t+s, 1).$$

Hence β has a "reaction time" of s, which means it anticipates if $s<0$. The set $\Gamma(s)$ of mappings from $N \to M$ is defined in the obvious analogous way.

LEMMA A1. *Let $I<\infty$, $\delta>0$, and $\beta \in \Delta$ be given such that*

$$\sup_{u \in M} C(u, \beta[u]) \leqq I. \tag{A.1}$$

Then there exists $\beta' \in \Delta$ and $C_1<\infty$ such that for all $u \in M$,

$$\left|\frac{d}{dt}\beta'[u](t)\right| \vee |\beta'[u](t)| \leqq C_1 \quad (a.s.), \tag{A.2}$$

$$C^{-\delta}(u, \beta'[u]) \leqq I. \tag{A.3}$$

(As before, $C^{-\delta}$ is the cost associated with the domain $G^{-\delta}$.) Furthermore, there exists $s<0$ such that given $\beta \in \Delta(s)$ satisfying (A.1) *there exists $\beta'' \in \Delta$ such that* (A.3) *holds for all $u \in M$ (with β'' replacing β' there.)*

Proof. The cost associated with β is simply $\frac{1}{2}\int_0^1 (\beta[u](t))^2\, dt \leqq I$, since exit before time $t=1$ *must* occur. Define

$$S(u, C_1) = \{t: |\beta[u](t)| \geqq C_1\},$$

$$\beta_1[u](t) = \begin{cases} 0, & t \in S(u, C_1), \\ \beta[u](t), & t \notin S(u, C_1). \end{cases}$$

Then β_1 is obviously a strategy, and

$$\frac{1}{2}\int_0^1 (\beta_1[u](t))^2\, dt \leqq \frac{1}{2}\int_0^1 (\beta[u](t))^2\, dt.$$

To show $C^{-\delta}(u, \beta_1[u]) \leqq C(u, \beta[u])$, it is sufficient to prove that if ϕ and ψ are defined by

$$\begin{aligned} \dot{\phi} &= b(\phi, u) + \sigma(\phi)\beta[u] \\ &= b(\phi, u) + \sigma(\phi)\beta_1[u] + \sigma(\phi)\beta[u] I_{S(u, C_1)}(t), \\ \dot{\psi} &= b(\psi, 0) + \sigma(\psi)\beta_1[u], \qquad \phi(0) = \psi(0) = x, \end{aligned}$$

then $d(\phi, \psi) \leqq \delta$. First note that

$$\left|\int_0^t \sigma(\phi(s))\beta[u](s) I_{S(u, C_1)}(s)\, ds\right| \leqq 2BI/C_1$$

for $0 \leqq t \leqq 1$. Hence,

$$|\phi(t) - \psi(t)| \leqq \int_0^t \bar{K}|\phi(s) - \psi(s)|\, ds + \int_0^t \bar{K}|\phi(s) - \psi(s)||\beta_1[u](s)|\, ds + 2BI/C_1.$$

Using the inequality $ab \leqq (a^2+b^2)/2$ in the second integral, and the Gronwall inequality, we obtain

$$d(\phi, \psi) \leqq 2BI[1 + \bar{K}(2+I)\, e^{\bar{K}(2+I)}]/C_1.$$

By choosing C_1 large, we have

$$C^{-\delta}(u, \beta'[u]) \leqq C(u, \beta[u])$$

for all $u \in M$.

Next we obtain β' by smoothing β_1. For $\Delta > 0$, define

$$\beta'[u](t) = \frac{1}{\Delta}\int_{t-\Delta}^{t} \beta_1[u](s)\, ds$$

(we define $\beta_1[u](s) = 0$ for $s < 0$). Obviously β' satisfies (A.2). We also have

$$\frac{1}{2}\int_0^1 (\beta'[u](t))^2\, dt \leqq \frac{1}{2}\int_0^1 (\beta_1[u](t))^2\, dt.$$

Formula (A.3) now follows if we can show that small $\Delta > 0$ implies that the solutions of

$$\dot{\phi} = b(\phi, u) + \sigma(\phi)\beta'[u], \qquad \phi(0) = x,$$
$$\dot{\psi} = b(\psi, u) + \sigma(\psi)\beta_1[u], \qquad \psi(0) = x$$

satisfy $d(\phi, \psi) \leqq \delta$. This follows from another application of Gronwall's lemma and an integration by parts.

Finally we consider the last statement of the lemma.

Let $s < 0$ be given. By the same argument as above we may assume the existence of $\beta' \in \Delta(s)$ satisfying (A.2) and (A.3). Define

$$\beta''[u](t) = \begin{cases} 0, & 0 \leqq t \leqq -s, \\ \beta'[u](t+s), & -s < t \leqq 1, \end{cases}$$
$$\dot{\phi} = b(\phi, u) + \sigma(\phi)\beta''[u], \qquad \phi(0) = x,$$
$$\dot{\psi} = b(\psi, u) + \sigma(\psi)\beta'[u], \qquad \psi(0) = x.$$

Then $\beta'' \in \Delta$. Arguments such as those used above, combined with the boundedness of β', β'' imply that when $s < 0$ is sufficiently large $d(\phi, \psi) \leqq \delta$. Hence we have $\beta'' \in \Delta$ such that

$$C^{-2\delta}(u, \beta''[u]) \leqq C(u, \beta[u]),$$

and the lemma is proved. □

LEMMA A2. *Let I, $\delta > 0$, and $\alpha \in \Gamma$ be given such that*

$$\inf_{v \in N} C(\alpha[v], v) \geqq I. \tag{A.4}$$

Then there exists $\alpha' \in \Gamma$ such that for all $v \in N$

$$\alpha'[v](\cdot) \text{ is a piecewise constant function,} \tag{A.5}$$

$$C^{\delta}(\alpha'[v], v) \geqq I. \tag{A.6}$$

Furthermore, there is an $s < 0$ such that given $\alpha \in \Gamma(s)$ satisfying (A.4) *there exists $\alpha'' \in \Gamma$ such that* (A.6) *holds for all $v \in N$ (where α'' replaces α' there).*

Proof. N may be written as the disjoint union $N = N_1 \cup N_2 \cup N_3$ with

$$N_1 = \{v \in N: \chi(\phi(\tau_x)) = 0\},$$
$$N_2 = \left\{ v \in N: \chi(\phi(\tau_x)) = \infty \text{ and } \frac{1}{2}\int_0^1 v^2\, dt \geqq I \right\},$$
$$N_3 = \left\{ v \in N: \chi(\phi(\tau_x)) = \infty \text{ and } \frac{1}{2}\int_0^1 v^2\, dt < I \right\}$$

(here $\dot{\phi} = b(\phi, \alpha[v]) + \sigma(\phi)v$, $\phi(0) = x$, and $\tau_x = \inf\{t: \phi(t) \in \partial G\} \wedge 1$). It is clear that we may define α' in any way we like on N_1 and N_2, as long as it is a strategy. For $\varepsilon > 0$ let $\{u_i, i = 1, \cdots, J\}$ be an ε-net of the control space K, and let $\{K_i, i = 1, \cdots, J\}$ be a Borel measurable partition of K such that the Hausdorff distance between $\{u_i\}$ and K_i is less than ε for $i = 1, \cdots, J$. For $\gamma > 0$, and $0 \leqq l \leqq 1/\gamma$, define

$$\tau(i, l, v) = \int_{l\gamma}^{l\gamma+\gamma} I_{\{\alpha[v](t) \in K_i\}}\, dt.$$

Then for all $v \in N$, l,

$$\sum_1^J \tau(i, l, v) = \gamma.$$

We define $\alpha'[v]$ by $\alpha'[v](t) = u_1$, for $0 \leqq t \leqq \gamma$, and

$$\alpha'[v](t) = u_i \quad \text{for } t \in \left(l\gamma + \sum_1^{i-1} \tau(j, l-1, v),\ l\gamma + \sum_1^{i} \tau(j, l-1, v) \right),$$

$l = 1, \cdots, 1/\gamma$.

Owing to the definition of α', we have

$$\sup_{0 \leqq t \leqq 1} \left| \int_0^t [b(\phi(r), \alpha[v](r)) - b(\phi(r), \alpha'[v](r))]\, dr \right| \leqq \varepsilon \bar{K} + \gamma B$$

for every $v \in N$ and measurable function $\phi(\cdot)$ taking values in G^δ. Define

$$\dot{\phi} = b(\phi, \alpha[v]) + \sigma(\phi)v, \qquad \phi(0) = x,$$

$$\dot{\psi} = b(\psi, \alpha'[v]) + \sigma(\psi)v, \qquad \psi(0) = x.$$

In order to prove (A.6) it is sufficient to prove $d(\phi, \psi) \leqq \delta$ when ε and γ are sufficiently small, and when $v \in N_3$. Using the estimate

$$\begin{aligned} |\phi(t) - \psi(t)| &\leqq \left| \int_0^t [b(\phi, \alpha[v]) - b(\phi, \alpha'[v])]\, ds \right| \\ &\quad + \left| \int_0^t [b(\phi, \alpha'[v]) - b(\psi, \alpha'[v]) + \sigma(\phi)v - \sigma(\psi)v]\, ds \right| \\ &\leqq \varepsilon \bar{K} + \gamma B + 3\bar{K} \int_0^t |\phi - \psi|\, ds/2 + \bar{K} \int_0^t |\phi - \psi| v^2\, ds/2, \end{aligned}$$

and Gronwall's lemma, we obtain

$$d(\phi, \psi) \leqq (\varepsilon \bar{K} + \gamma B)[1 + \bar{K}(2 + I) \exp \bar{K}(2 + I)]. \tag{A.7}$$

Hence we obtain (A.6) for small ε, γ.

If we are given $\alpha \in \Gamma(s)$, and define

$$\alpha''[v](t) = \begin{cases} u_1 & \text{for } 0 \leqq t \leqq -s, \\ \alpha[v](t+s) & \text{for } -s < t \leqq 1, \end{cases}$$

then $\alpha'' \in \Gamma$, and by the same argument as above we can obtain (A.6) when $s < 0$ is sufficiently large. The only difference is that in (A.7) we replace $\varepsilon \bar{K} + \gamma B$ by $-sB$. □

LEMMA A3. *Given $\rho > 0$ and $M < \infty$, there exist $\Delta_0 > 0$ and $\varepsilon_0 > 0$ such that* (3.13) *holds for $\varepsilon \leqq \varepsilon_0$, $\Delta \leqq \Delta_0$.*

Proof. We begin by defining a stopping time (all stopping times are with respect to $w(\cdot)$) for $\rho_1 > 0$:

$$\tau_1 = \inf\{t : |x^{\varepsilon,\Delta}(t) - x^{\varepsilon,\Delta}([t/\Delta]\Delta)| \geqq \rho_1\} \wedge 1.$$

A simple calculation shows there exist $\varepsilon_{0,1} > 0$ and $\Delta_{0,1} > 0$ (depending on ρ_1) such that $\varepsilon \leqq \varepsilon_{0,1}$ and $\Delta \leqq \Delta_{0,1}$ imply

$$P_x\{\tau_1 < 1\} \leqq \exp(-(M+2)/\varepsilon^2). \tag{A.8}$$

Next we rewrite the equation for $x^{\varepsilon,\Delta}$ as

$$dx^{\varepsilon,\Delta} = b(x^{\varepsilon,\Delta}, u)\, dt + \varepsilon\sigma(x^{\varepsilon,\Delta})\, dw + d\gamma^{\varepsilon,\Delta}, \qquad x^{\varepsilon,\Delta}(0) = x, \tag{A.9}$$

where

$$\begin{aligned} &d\gamma^{\varepsilon,\Delta} = \varepsilon\sigma(x^{\varepsilon,\Delta})[v^\Delta\, dt - dw], \\ &\int_{i\Delta}^{i\Delta+\Delta} \varepsilon\sigma(x^{\varepsilon,\Delta}(i\Delta))v^\Delta(t)\, dt = \int_{i\Delta-\Delta}^{i\Delta} \varepsilon\sigma(x^{\varepsilon,\Delta}(i\Delta))\, dw(t). \end{aligned} \tag{A.10}$$

We therefore have decomposition

$$\gamma^{\varepsilon,\Delta}(t) = I_1(t) + I_2(t) + I_3(t) + I_4(t),$$

where (for $k = [t/\Delta] - 1$)

$$\begin{aligned} I_1(t) &= -\sum_1^k \int_{i\Delta-\Delta}^{i\Delta} \varepsilon[\sigma(x^{\varepsilon,\Delta}(s)) - \sigma(x^{\varepsilon,\Delta}(i\Delta))]\, dw(s), \\ I_2(t) &= \sum_1^k \int_{i\Delta}^{i\Delta+\Delta} \varepsilon[\sigma(x^{\varepsilon,\Delta}(s)) - \sigma(x^{\varepsilon,\Delta}(i\Delta))]v^\Delta(s)\, ds, \\ I_3(t) &= \int_{k\Delta+\Delta}^{t} \varepsilon\sigma(x^{\varepsilon,\Delta}(s))v^\Delta(s)\, ds, \qquad I_4(t) = \int_{k\Delta}^{t} \varepsilon\sigma(x^{\varepsilon,\Delta}(s))\, dw(s). \end{aligned}$$

For $\rho_2 > 0$, define the stopping times

$$\tau_{2,i} = \inf\{t : |I_i(t)| \geqq \rho_2/4\} \wedge 1.$$

The same estimates as those used to show (A.8) give the existence of $0 < \varepsilon_{0,2} \leqq \varepsilon_{0,1}$, and $0 < \Delta_{0,2} \leqq \Delta_{0,1}$ such that for $\varepsilon \leqq \varepsilon_{0,2}$ and $\Delta \leqq \Delta_{0,2}$,

$$P_x\{\tau_{2,i} < 1\} \leqq \exp -(M+1)/\varepsilon^2$$

for $i = 3, 4$.

Next consider $\tau_{2,1}$. Using

$$P_x\{\tau_{2,1} < 1\} \leqq P_x\{\tau_{2,1} < 1,\ \tau_1 = 1\} + P_x\{\tau_1 < 1\},$$

(A.8), and a standard estimate on stochastic integrals [8, Lemma 4.7], by picking ρ_1 small we obtain $0 < \varepsilon'_{0,2} \leqq \varepsilon_{0,2}$ and $0 < \Delta'_{0,2} \leqq \Delta_{0,2}$ such that $\varepsilon \leqq \varepsilon'_{0,2}$ and $\Delta \leqq \Delta'_{0,2}$ imply

$$P_x\{\tau_{2,1} < 1\} \leqq \exp -(M+1)/\varepsilon^2.$$

Finally we consider $\tau_{2,2}$. Using the Lipschitz property of $\sigma(\cdot)$, we have the following bound on a typical summand in $I_2(t)$:

$$\begin{aligned} &\int_{i\Delta}^{i\Delta+\Delta} \varepsilon[\sigma(x^{\varepsilon,\Delta}(s)) - \sigma(x^{\varepsilon,\Delta}(i\Delta))]v^\Delta(s)\, ds \\ &\quad \leqq \int_{i\Delta}^{i\Delta+\Delta} \varepsilon\bar{K} \left| \int_{i\Delta}^{t} (b(x^{\varepsilon,\Delta}(s)) + \varepsilon\sigma(x^{\varepsilon,\Delta}(s))v^\Delta(s))\, ds \right| |v^\Delta(t)|\, dt \\ &\quad \leqq \varepsilon\bar{K}B\Delta^2|v^\Delta(i\Delta)|/2 + \varepsilon^2\bar{K}B\Delta^2|v^\Delta(i\Delta)|^2/2. \end{aligned}$$

We therefore have

$$P_x\{\tau_{2,2}<1\} \leqq P\left\{\varepsilon\Delta^2\bar{K}B\sum_1^{1/\Delta}|\theta_i| \geqq \frac{\rho_2}{4}\right\} + P\left\{\varepsilon^2\Delta^2\bar{K}B\sum_1^{1/\Delta}|\theta_i|^2 \geqq \frac{\rho_2}{4}\right\}, \tag{A.11}$$

where $\{\theta_i\}$ is a sequence of independent and identically distributed $N(0, 1/\Delta)$ random variables. For the sake of notational simplicity, we estimate these terms in the case where $\{\theta_i\}$ is a scalar-valued sequence.

Using $E \exp c\theta_i^2 = (1-2c/\Delta)$ (for $2c/\Delta<1$), we obtain (for any $\xi>0$ such that $2\varepsilon^2\Delta\bar{K}B\xi<1$)

$$P\left\{\varepsilon^2\Delta^2\bar{K}B\sum_1^{1/\Delta}|\theta_i|^2 \geqq \frac{\rho_2}{4}\right\} \leqq \left(\exp -\frac{\xi\rho_2}{4}\right)(1-2\varepsilon^2\Delta\bar{K}B\xi)^{1/\Delta}$$

$$= \exp\left[-\frac{\xi\rho_2}{4} + \frac{1}{\Delta}\log(1-2\varepsilon^2\Delta\bar{K}B\xi)\right].$$

Now take $\xi = (M+2)4/\rho_2\varepsilon^2$, and use the fact that the log term $\to -8(M+2)\bar{K}B/\rho_2$ as $\Delta\to 0$ to get the estimate of the type (A.8) for the second term of (A.11).

For the first term of (A.11), we will use the fact that $E \exp c|\theta_i| \leqq 2E \exp c\theta_i = 2\exp c^2/2\Delta$. For $\xi>0$ we have

$$P\left\{\varepsilon\Delta^2\bar{K}B\sum_1^{1/\Delta}|\theta_i| \geqq \frac{\rho_2}{4}\right\} \leqq \exp -\frac{\xi\rho_2}{4}\cdot\exp\xi^2\varepsilon^2\Delta^2\bar{K}^2B^2/2\cdot\exp\frac{1}{\Delta}\log 2.$$

Minimizing with respect to $\xi>0$, we obtain the bound

$$\exp[-\rho_2^2/32\varepsilon^2\Delta^2\bar{K}^2B^2 + (\log 2)/\Delta],$$

which again gives the desired bound of type (A.8) for small Δ, ε.

Hence there are $0<\varepsilon''_{0,2} \leqq \varepsilon'_{0,2}$ and $0<\Delta''_{0,2} \leqq \Delta'_{0,2}$ such that for $\varepsilon \leqq \varepsilon''_{0,2}$ and $\Delta \leqq \Delta''_{0,2}$,

$$P_x\{\tau_{2,2}<1\} \leqq \exp -(M+1)/\varepsilon^2.$$

Now set $\tau_2 = \bigwedge_1^4 \tau_{2,i}$. On the set where $\tau_2 = 1$, $\sup_{0\leqq t\leqq 1}|\gamma^{\varepsilon,\Delta}(t)| \leqq \rho_2$. We have shown that for ε sufficiently small, $P_x\{\tau_2<1\} \leqq \exp -(M+1)/\varepsilon^2$. These facts, together with a standard estimate in large deviations theory [2, Proof of Lemma 6.2], yield the lemma. □

LEMMA A4. *Assume* A1 *and* A2. *Then* $I^-(x) \leqq I^+(x)$.

Proof. Let $c>0$ be given. By A2 there is $\delta>0$ such that $I^-(x) \leqq I^-(x,-\delta)+c$, $I^+(x) \geqq I^+(x,\delta)-c$. Next choose $s<0$ such that the second statements of Lemmas A1 and A2 hold, with $I^-(x)+1$ (respectively, $I^+(x)-1$) replacing I in Lemma A1 (respectively, A2). Suppose $\beta\in\Delta(s)$ is a c-optimal solution to the problem

$$\inf_{\beta\in\Delta(s)}\sup_{u\in M} C(u,\beta[u]). \tag{A.12}$$

Let $\bar{I}^-(s)$ denote the value of the expression given in (A.12). Then by Lemma A1 we may find $\beta''\in\Delta$ such that

$$\sup_{u\in M} C^{-\delta}(u,\beta''[u]) \leqq \bar{I}^-(s).$$

Hence we may conclude $I^-(x,-\delta) \leqq \bar{I}^-(s)$. In an analogous manner we may prove $I^+(x,\delta) \geqq \bar{I}^+(s)$, where

$$\bar{I}^+(s) = \sup_{\alpha\in\Gamma(s)}\inf_{v\in N} C(\alpha[v],v).$$

It follows that $I^-(x) - I^+(x) \leqq \bar{I}^-(s) - \bar{I}^+(s) + 2c$. Since $c > 0$ is arbitrary, we are finished if we can show there is $s_0 < 0$ such that $\bar{I}^-(s) \leqq \bar{I}^+(s)$ for all $s_0 < s < 0$. However, as is proved in [4, p. 17], when $-2^{-N} < s$, $\bar{I}^-(s)$ is a lower bound for the value v_N^- defined in the sense of Friedman having stepsize 2^{-N} and allowing the minimizing player to move first (for the full definition of values in the sense of Friedman, see [4, § 3]). An analogous statement holds for the corresponding upper values; $v_N^+ \leqq \bar{I}^+(s)$. Since (as is easily proved) $v_N^- \leqq v_N^+$ for every N [4, p. 11], we are finished. □

REFERENCES

[1] A. FRIEDMAN, *Differential Games*, CBMS-NSF Regional Conference Series in Mathematics, American Mathematical Society, Providence, RI, 1974.

[2] S. R. S. VARADHAN, *Large Deviations and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics 46, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984.

[3] W. H. FLEMING AND P. E. SOUGANIDIS, PDE-*viscosity solution approach to some problems of large deviations*, Ann. Scuol. Norm. Sup. Pisa Cl. Sci. 4, 13 (1986), pp. 171-192.

[4] R. J. ELLIOTT AND M. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).

[5] A. D. WENTZELL AND M. I. FREIDLIN, *Some problems concerning stability under small random perturbations*, Theory Probab. Appl., 17 (1972), pp. 269-283.

[6] S. M. MEERKOV AND T. RUNOLFSSON, *Aiming control*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 494-498.

[7] K. L. CHUNG, *Lectures from Markov Processes to Brownian Motion*, Grundlehren Math. Wiss. 249, Springer-Verlag, Berlin, New York, 1982.

[8] D. STROOK, *An Introduction to the Theory of Large Deviations*, Springer-Verlag, Berlin, New York, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 2, pp. 446-455, March 1989

011

OPTIMAL CONTROL OF SEMILINEAR MULTISTATE SYSTEMS WITH STATE CONSTRAINTS*

JOSEPH FREDERIC BONNANS† AND EDUARDO CASAS‡

Abstract. This paper deals with state-constrained optimal control problems governed by a semilinear multistate equation. The authors prove the existence of solutions and derive optimality conditions.

Key words. optimal control, subdifferential calculus, optimality conditions, elliptic operators, semilinear equations, multistate systems

AMS(MOS) subject classifications. 49B22, 49A22

1. Introduction. This paper is concerned with state-constrained optimal control problems governed by a semilinear elliptic operator. As we make no monotonicity assumption, the state equation may be unsolvable or may have several solutions. These kinds of ill-posed systems may arise in connection with bifurcation theory; some models arising in enzymatic reactions, plasma physics, and chemistry have this property (see some examples in Crandall and Rabinowitz [11] and Lions [15]). However, this paper studies only a model problem. Our aim is to obtain existence results and to derive the optimality system.

There exists a vast literature on the control of well-posed state-constrained systems. The subdifferential calculus of convex analysis is a useful tool for dealing with linear state equations (see Mackenroth [16], [17], Bonnans and Casas [7], and Casas [8], [9]). In the nonlinear case, Bonnans and Casas [4]-[6] derived the optimality system using the results of Clarke [10].

The control of nonmonotone elliptic systems, but without state constraints, has been studied by Lions [15] (see also Komornik [14]). The optimality system is derived there by penalizing the state equation and passing to the limit in the optimality conditions of the penalized problem.

The novelty of this paper lies in the simultaneous presence of state constraints and of an ill-posed system. Our method consists of approximating the problem by removing the nonlinearity from the state equation and penalizing a part of the state constraints. We formulate the problem and obtain an existence result in § 2, derive the optimality system in § 3, and study several examples in § 4.

2. Formulation of the control problem. Let Ω be an open bounded subset of $\mathbb{R}^n$ ($n \leqq 3$) with C^2 boundary Γ. Let us consider the following system:

$$\begin{aligned} &Ay + \phi(y) = u \quad \text{in } \Omega, \\ &y = 0 \quad \text{on } \Gamma, \end{aligned} \tag{2.1}$$

where

$$Ay = -\sum_{i,j=1}^{n} \partial_{x_j}(a_{ij}(x)\partial_{x_i} y) + a_0(x)y,$$

* Received by the editors August 3, 1987; accepted for publication (in revised form) April 1, 1988.

† Institut National de Recherche en Informatique et en Automatique, Rocquencourt, 78153 Le Chesnay, France.

‡ Departamento de Mátematicàs, Estadística y Computacion, Facultad de Ciencias, 39005 Santander, Spain. The work of this author was supported in part by Dirección general de investigación científica y técnica.

$$
\begin{aligned}
&a_0 \in L^\infty(\Omega), \qquad a_0(x) \geqq 0 \quad \text{a.e. } x \in \Omega,\\
&a_{ij} \quad \text{is Lipschitz on } \bar{\Omega}\ (1 \leqq i, j \leqq n),\\
&\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq \alpha_0 \|\xi\|^2, \qquad \alpha_0 > 0 \quad \forall \xi \in R^n, \quad \forall x \in \Omega,
\end{aligned} \tag{2.2}
$$

$$
\phi : R \to R \quad \text{is } C^1. \tag{2.3}
$$

Let K be a nonempty, convex, closed subset of $L^2(\Omega)$, σ be greater than or equal to 2, N be nonnegative, and y_d in $L^\sigma(\Omega)$ be given, and let $J : L^\sigma(\Omega) \times L^2(\Omega) \to \mathbb{R}$ be the functional

$$
J(y, u) = \frac{1}{\sigma} \int_\Omega |y(x) - y_d(x)|^\sigma \, dx + \frac{N}{2} \int_\Omega u^2(x) \, dx. \tag{2.4}
$$

Let Z be a separable Banach space, B be a closed convex subset of Z with nonempty interior, and a be given in $\mathbb{R}^m$ ($m \geqq 0$; we identify $\mathbb{R}^0$ and $\{0\}$). Define $Y = H^2(\Omega) \cap H_0^1(\Omega)$, where $H^s(\Omega)$ and $H_0^s(\Omega)$ are the usual Sobolev spaces (see Adams [1], Nečas [18]). Let $C_0(\Omega)$ be the space of real continuous functions on $\bar{\Omega}$ vanishing on Γ, endowed with the supremum norm $\| \ \|_\infty$. It is known that Y is compactly embedded in $C_0(\Omega)$ for $n \leqq 3$. The dual of $C_0(\Omega)$ is the space $M(\Omega)$ of real and regular Borel measures on Ω, endowed with the norm

$$
\|\mu\|_{M(\Omega)} = |\mu|(\Omega),
$$

where $|\mu|$ is the total variation measure of μ (Rudin [19]). Finally, let $T : C_0(\Omega) \to \mathbb{R}^m$ and $L : C_0(\Omega) \to Z$ be linear continuous mappings. In order to derive the optimality conditions, we will suppose that

$$
T(Y) = \mathbb{R}^m \quad \text{and} \quad \overline{L(Y)} = Z. \tag{2.5}
$$

We consider the following control problem:

$$
\begin{aligned}
&\text{minimize} \quad J(y, u)\\
&\text{subject to} \quad (2.1), \quad u \in K, \quad y \in Y, \quad Ty = a, \quad Ly \in B.
\end{aligned} \tag{P}
$$

Remark 1. The assumptions on Ω and A imply (Nečas [18]) that for each f in $L^2(\Omega)$ there exists a unique solution $y \in Y$ of the Dirichlet problem

$$
Ay = f \quad \text{in } \Omega, \qquad y = 0 \quad \text{on } \Gamma,
$$

and moreover, there exists C_1 independent of f such that

$$
\|y\|_{H^2(\Omega)} \leqq C_1 \|f\|_{L^2(\Omega)}. \tag{2.6}
$$

In fact all our results still hold if we assume that Ω is bounded, Y is compactly embedded in $C_0(\Omega)$, and (2.6) holds. This is the case, for instance, if A is symmetric and satisfies (2.2) and Ω is bounded and convex (Grisvard [13]).

Remark 2. The existence of several states associated to the same control has been obtained, e.g., with cubic nonlinearities [11]. The inclusion of Y in $C_0(\Omega)$ for $n \leqq 3$ (Adams [1]) implies that $A + \phi$ maps Y into $L^2(\Omega)$; hence all elements of Y are associated to a control. For parabolic systems the situation is essentially different (Bonnans [3]).

Let us now give some examples of control problem that fall into the previous formulation.

$$\text{(P1)} \qquad \begin{aligned} &\text{minimize} \quad J(y,u) \\ &\text{subject to} \quad (2.1), \quad u \in K, \quad y \in Y, \quad y(x_i) = a_i, \quad 1 \leqq i \leqq m. \end{aligned}$$

Here $\{x_i\}$ are given in Ω and we may take $B = Z = C_0(\Omega)$, L is the identity in $C_0(\Omega)$, and $Ty = \{y(x_i)\}$.

$$\text{(P2)} \qquad \begin{aligned} &\text{minimize} \quad J(y,u) \\ &\text{subject to} \quad (2.1), \quad u \in K, \quad y \in Y, \quad \int_\Omega |y(x)|\,dx \leqq \delta \quad \text{with } \delta > 0. \end{aligned}$$

Here $m = 0$, $T = 0$, $Z = L^1(\Omega)$, B is the closed ball with center zero and radius δ, and L is the canonical injection from $C_0(\Omega)$ into $L^1(\Omega)$.

$$\text{(P3)} \qquad \begin{aligned} &\text{minimize} \quad J(y,u) \\ &\text{subject to} \quad (2.1), \quad u \in K, \quad y \in Y, \quad \int_\Omega y(x)\,dx = a, \\ &|y(x)| \leqq \delta \quad \forall x \in \Omega, \quad \text{with } \delta > 0. \end{aligned}$$

Here $m = 1$ and $Ty = \int_\Omega y(x)\,dx$, $Z = C_0(\Omega)$, B is the closed ball with radius δ and center zero, and L is the identity. These three examples obviously satisfy (2.5).

We now give a result concerning the existence of a solution to problem (P). For this we need a relation between σ and the nonmonotone part of ϕ.

THEOREM 1. *Suppose* (2.2) *and* (2.3) *hold, and suppose the following*:

(i) *There exists* (y, u) *satisfying the constraints of* (P) (*i.e.*, (P) *is feasible*).

(ii) *Either* $N > 0$ *or* K *is bounded in* $L^2(\Omega)$.

(iii) *We may write* $\phi(t) = \phi_1(t) + \phi_2(t)$, *with* ϕ_i *continuous,* $i = 1, 2$, $\phi_1(t)$ *nondecreasing, and such that for some* $C > 0$

$$|\phi_2(t)| \leqq C(1 + |t|^{\sigma/2}).$$

Then problem (P) *has* (*at least*) *one solution.*

Proof. As (P) is feasible, there exists a minimizing sequence $\{(y_n, u_n)\}$ in $Y \times K$. Because of (ii), $\{u_n\}$ is bounded in $L^2(\Omega)$. We are going to prove that $\{Ay_n\}$ is bounded in $L^2(\Omega)$, and for this we may assume that ϕ_1 is differentiable. Otherwise, we would approximate ϕ_1 by a standard convolution technique and then pass to the limit. We also may assume without loss of generality that $\phi_1(0) = 0$.

The form of J implies that $\{y_n\}$ is bounded in $L^\sigma(\Omega)$; hence with (iii), $\phi_2(y_n)$ is bounded in $L^2(\Omega)$, as is $f_n = -\phi_2(y_n) + u_n = Ay_n + \phi_1(y_n)$. As $\phi_1(y_n)$ is in $C_0(\Omega)$, Ay_n belongs to $L^2(\Omega)$. Computing the scalar product of f_n with Ay_n in $L^2(\Omega)$, and integrating the nonlinear term by parts, we obtain

$$\|Ay_n\|^2_{L^2(\Omega)} + \int_\Omega \phi_1'(y_n) \sum_{i,j=1}^{n} a_{ij}(x) \frac{\partial y_n}{\partial x_i} \frac{\partial y_n}{\partial x_j}\,dx$$

$$+ \int_\Omega a_0(x) \phi_1(y_n(x)) y_n(x)\,dx \leqq \|f_n\|_{L^2(\Omega)} \|Ay_n\|_{L^2(\Omega)}.$$

The second and third term of the left-hand side are nonnegative because of (2.2), the monotonicity of ϕ_1, and the equality $\Phi_1(0) = 0$. Hence $\|Ay_n\|$ is bounded in $L^2(\Omega)$;

with (2.6), this implies that $\{y_n\}$ is bounded in Y. As Y is compactly embedded in $C_0(\Omega)$ for $n \leqq 3$, selecting a subsequence if necessary, we may assume that

$$y_n \to \bar{y} \quad \text{weakly in } Y, \quad \text{strongly in } C_0(\Omega),$$
$$Ay_n \to A\bar{y} \quad \text{weakly in } L^2(\Omega),$$
$$u_n \to \bar{u} \quad \text{weakly in } L^2(\Omega).$$

This implies $T\bar{y} = a$, $L\bar{y} \in B$, and $\phi(y_n) \to \phi(\bar{y})$ in $C_0(\Omega)$; hence Ay_n weakly converges in $L^2(\Omega)$ toward $\bar{u} - \phi(\bar{y})$; hence $(\bar{y}, \bar{u})$ satisfies (2.1). As K is closed and convex, hence weakly closed, $\bar{u}$ is in K. Finally, the convexity and continuity of J implies its weak lower semicontinuity; the result follows. □

3. The optimality system. For any set C, denote its indicatrix by I_C, defined by

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

We denote the subdifferential of a convex function f by ∂f (see Barbu and Precupanu [2], Ekeland and Temam [12]). The spaces $W_0^{1,s}(\Omega)$ and $W^{1,s}(\Omega)$ are the usual Sobolev spaces (Adams [1]). We denote by T^* the adjoint operator of T and by $R(T^*)$ its range. The aim of this section is to prove the following result.

THEOREM 2. *Let $(\bar{y}, \bar{u})$ be a solution of* (P). *We assume that* (2.2)-(2.5) *hold and that*

$$\partial(I_B \circ L)(\bar{y}) \cap R(T^*) = \{0\}. \tag{3.1}$$

Then there exists $\bar{p}$ in $W_0^{1,s}(\Omega)$ for all $s < n/(n-1)$, $\bar{\lambda}$ in $\mathbb{R}^m$, $\bar{\mu}$ in Z', and $\bar{\alpha} \geqq 0$ such that

$$\bar{\alpha} + \|\bar{p}\|_{W_0^{1,s}(\Omega)} > 0, \tag{3.2}$$

$$A^*\bar{p} + \phi'(\bar{y})\bar{p} = \bar{\alpha}|\bar{y} - y_d|^{\sigma-2}(\bar{y} - y_d) + T^*\bar{\lambda} + L^*\bar{\mu}, \tag{3.3}$$

$$\langle \bar{\mu}, z - L\bar{y} \rangle \leqq 0 \quad \forall z \in B, \tag{3.4}$$

$$\int_\Omega (\bar{p} + \bar{\alpha} N\bar{u})(v - \bar{u})\, dx \geqq 0 \quad \forall v \in K. \tag{3.5}$$

Remark 3. Since B has a nonempty interior, we deduce from (2.5) that $R(L) \cap \mathring{B} \neq \varnothing$. This implies (see Barbu and Precupanu [2], Ekeland and Teman [12]) that $\partial(I_B \circ L) \cdot (\bar{y}) = L^* \partial I_B(L\bar{y})$.

Remark 4. We will verify that hypothesis (3.1) holds in our three examples. However, if (3.1) does not hold, then by Remark 3 there exists $(\bar{\lambda}, \bar{\mu})$ in $\mathbb{R}^m \times \partial I_B(L\bar{y})$ such that $\|\bar{\lambda}\| + \|\bar{\mu}\| > 0$ and $T^*\bar{\lambda} + L^*\bar{\mu} = 0$. In other words, if all hypotheses of Theorem 2 are satisfied except perhaps (3.1), there exist $\bar{p}$, $\bar{\lambda}$, $\bar{\mu}$, $\bar{\alpha}$ as in Theorem 1, not all null, satisfying (3.3)-(3.5).

In order to prove Theorem 2, we need to establish some preliminary results.

LEMMA 1. *Let W be a Banach space and D be a convex subset of W (not necessarily closed) with nonempty interior. Let $\{(w_n, \eta_n)\}$ be a sequence in $W \times W'$ such that $w_n \in D$, $w_n \to w$ and $\eta_n \in \partial I_D(w_n)$. If $\liminf \|\eta_n\| > 0$, then zero is not a weak-star limit point of $\{\eta_n\}$.*

Proof. Assume that the conclusion does not hold. Let w_0 be given in $\mathring{D}$. There exists $r > 0$ such that $\|w\| \leqq r$ implies that $w_0 + w$ is in D; hence

$$\langle \eta_n, w_0 + w - w_n \rangle \leqq 0,$$

and this implies

$$r\|\eta_n\| = \sup_{\|w\| \leqq r} \langle \eta_n, w \rangle \leqq \langle \eta_n, w_n - w_0 \rangle.$$

The strong convergence of w_n allows us to pass to the limit and we get

$$r \liminf \|\eta_n\| \leqq 0,$$

which gives a contradiction. □

LEMMA 2. *Let W be a Banach space, and f (respectively, g) be a Gâteaux-differentiable (respectively, convex) mapping from W into $\mathbb{R}$ (respectively, $]-\infty, +\infty]$). Let $\bar{x}$ be a solution of the following problem*:

$$\min f(x)+g(x), \qquad x \in W.$$

Then

$$\langle \nabla f(\bar{x}), x-\bar{x}\rangle + g(x) - g(\bar{x}) \geqq 0 \quad \forall x \in W,$$

or, equivalently,

$$\nabla f(\bar{x}) + \partial g(\bar{x}) \ni 0.$$

Proof. A straightforward application of the definition of the subdifferential [12] allows us to verify the equivalence of the two statements of the conclusion. Now consider $x^t = \bar{x} + t(x-\bar{x})$ for t in $]0, 1[$. We have, using the convexity of g: $f(x^t) + g(x^t) \leqq f(x^t) + (1-t)g(\bar{x}) + tg(x)$; hence, as $\bar{x}$ is a solution of the problem above,

$$0 \leqq f(x^t) + g(x^t) - (f(\bar{x}) + g(\bar{x})) \leqq f(x^t) - f(\bar{x}) + t(g(x) - g(\bar{x})).$$

Dividing by t and passing to the limit, we obtain the result. □

We now consider the following approximate problem. Let the state equation be

$$\begin{aligned} Ay &= u + w \quad \text{in } \Omega, \\ y &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{3.6}$$

The control is now (u, w) in $L^2(\Omega) \times L^2(\Omega)$. We define

$$J_\varepsilon(y, u, w) = J(y, u) + \frac{1}{2\varepsilon} \int_\Omega (w + \phi(y))^2 \, dx + \frac{1}{2\varepsilon} \|Ty - a\|^2 + \frac{1}{2} \int_\Omega (u - \bar{u})^2 \, dx + \frac{1}{2} \int_\Omega (w + \phi(\bar{y}))^2 \, dx.$$

The approximate problem is

$$(\mathrm{P}_\varepsilon) \qquad \begin{aligned} &\text{minimize} \quad J_\varepsilon(y, u, w) \\ &\text{subject to} \quad (3.6), \quad u \in K, \quad w \in L^2(\Omega), \quad y \in Y, \quad Ly \in B. \end{aligned}$$

THEOREM 3. *Let $(\bar{y}, \bar{u})$ be a solution of* (P). *We assume that* (2.2)-(2.5) *hold. Then we have the following*:

(i) *Problem* (P_ε) *has at least one solution.*

(ii) *To each solution $(y_\varepsilon, u_\varepsilon, w_\varepsilon)$ of (P_ε) is associated p_ε in $W_0^{1,s}(\Omega)$ for all $s < n/(n-1)$, $\mu_\varepsilon \in Z'$, and λ_ε in $\mathbb{R}^m$ such that*

$$A^* p_\varepsilon = |y_\varepsilon - y_d|^{\sigma-2}(y_\varepsilon - y_d) + T^* \lambda_\varepsilon + L^* \mu_\varepsilon + \frac{1}{\varepsilon} \phi'(y_\varepsilon)(w_\varepsilon + \phi(y_\varepsilon)),$$

$$p_\varepsilon = 0 \quad \text{on } \Gamma,$$

$$\langle \mu_\varepsilon, z - Ly_\varepsilon \rangle \leqq 0 \quad \forall z \in B,$$

$$\int_\Omega (p_\varepsilon + Nu_\varepsilon + u_\varepsilon - \bar{u})(v - u_\varepsilon) \, dx \geqq 0 \quad \forall v \in K,$$

$$p_\varepsilon + \frac{1}{\varepsilon}[w_\varepsilon + \phi(y_\varepsilon)] + w_\varepsilon + \phi(\bar{y}) = 0.$$

Proof. (i) The triple $(\bar{y}, \bar{u}, -\phi(\bar{y}))$ is feasible for (P_ε). Any minimizing sequence is bounded in $L^\sigma(\Omega) \times L^2(\Omega) \times L^2(\Omega)$, and hence by (3.6) in $Y \times L^2(\Omega) \times L^2(\Omega)$. Taking a subsequence if necessary and using the compactness of Y in $C_0(\Omega)$ $(n \leqq 3)$ to pass to the limit in the nonlinear terms, we get the result as in the proof of Theorem 1.

(ii) Denote by $y_{u,w}$ the solution of (3.6) and by $\theta(u, w)$ the mapping $(u, w) \to J_\varepsilon(y_{u,w}, u, w)$. It is easy to verify that θ is C^1 and that

$$\theta'_u(u, w) = q + Nu + u - \bar{u},$$

$$\theta'_w(u, w) = q + \frac{1}{\varepsilon}(w + \phi(y_{u,w})) + w + \phi(\bar{y}),$$

where q is the solution of (A^* being the formal transpose of A):

$$A^*q = |y_{u,w} - y_d|^{\sigma-2}(y_{u,w} - y_d) + \frac{1}{\varepsilon}\phi'(y_{u,w})(w + \phi(y_{u,w})) + \frac{1}{\varepsilon}T^*(Ty_{u,w} - a) \quad \text{in } \Omega,$$

$$q = 0 \quad \text{on } \Gamma.$$

Let $(y_\varepsilon, u_\varepsilon, w_\varepsilon)$ be a solution of (P_ε) and q_ε the associated adjoint-state. Let us define

$$\hat{L}: L^2(\Omega) \times L^2(\Omega) \to Z,$$

$$(u, w) \to Ly_{u,w},$$

$$\hat{K} = K \times L^2(\Omega),$$

$$g(u, w) = I_B(\hat{L}(u,w)) + I_{\hat{K}}(u, w).$$

Problem (P_ε) is equivalent to

$$\min \theta(u, w) + g(u, w), (u, w) \in L^2(\Omega) \times L^2(\Omega).$$

Now applying Lemma 2, we get

$$\nabla\theta(u_\varepsilon, w_\varepsilon) + \partial g(u_\varepsilon, w_\varepsilon) \ni 0.$$

The mapping $w \to y_{u,w}$(with u fixed) is an isomorphism from $L^2(\Omega)$ onto Y. Hence by (2.1) there exists (u, w) in $\hat{K}$ with $\hat{L}(u, w)$ in $\mathring{B}$. This allows us [12] to apply the rules of subdifferential calculus to the mapping g and we get the equality

$$\partial g(u_\varepsilon, w_\varepsilon) = \hat{L}^* \partial I_B(Ly_\varepsilon) + \partial I_{\hat{K}}(u_\varepsilon, w_\varepsilon).$$

Hence there exists μ_ε in $\partial I_B(Ly_\varepsilon)$ such that

$$\nabla\theta(u_\varepsilon, w_\varepsilon) - \hat{L}^*\mu_\varepsilon + \partial I_{\hat{K}}(u_\varepsilon, w_\varepsilon) \ni 0,$$

or equivalently,

$$(\theta'_u(u_\varepsilon, w_\varepsilon), u - u_\varepsilon) + (\theta'_w(u_\varepsilon, w_\varepsilon), w - w_\varepsilon) + \langle \mu_\varepsilon, Ly_{u,w} - Ly_\varepsilon \rangle \geqq 0 \quad \forall (u, w) \in K \times L^2(\Omega).$$

Let r_ε be the solution of

$$A^*r_\varepsilon = L^*\mu_\varepsilon \quad \text{in } \Omega,$$

$$r_\varepsilon = 0 \quad \text{on } \Gamma.$$

We get

$$(\theta'_u(u_\varepsilon, w_\varepsilon) + r_\varepsilon, u - u_\varepsilon) \geqq 0 \quad \forall u \in K,$$

$$\theta'_w(u_\varepsilon, w_\varepsilon) + r_\varepsilon = 0.$$

We obtain the result with $p_\varepsilon = q_\varepsilon + r_\varepsilon$ and $\lambda_\varepsilon = (1/\varepsilon)(Ty_\varepsilon - a)$. As $A^* p_\varepsilon$ is in $M(\Omega)$, p_ε is in $W_0^{1,s}(\Omega)$ for all $s < n/(n-1)$ (see [9], [21]). □

LEMMA 3. *Let* $\{(y_\varepsilon, u_\varepsilon, w_\varepsilon)\}$ *be a sequence of solutions of* (P_ε). *Then*

$$0 = \lim_{\varepsilon \downarrow 0} \|y_\varepsilon - \bar{y}\|_Y = \lim_{\varepsilon \downarrow 0} \|u_\varepsilon - \bar{u}\|_{L^2(\Omega)} = \lim_{\varepsilon \downarrow 0} \|w_\varepsilon + \phi(\bar{y})\|_{L^2(\Omega)}.$$

Proof. From the inequality $J_\varepsilon(y_\varepsilon, u_\varepsilon, w_\varepsilon) \leqq J_\varepsilon(\bar{y}, \bar{u}, -\phi(\bar{y})) = J(\bar{y}, \bar{u})$ and the form of J, we deduce that $\{(y_\varepsilon, u_\varepsilon, w_\varepsilon)$ is bounded in $L^\sigma(\Omega) \times L^2(\Omega) \times L^2(\Omega)$; hence $\{y_\varepsilon\}$ is bounded in Y by (3.6) and (2.6). This implies that for $\varepsilon \in D$, D being a subset of $]0, \infty[$ having zero as limit point, we have for some (y, u, w) in $Y \times L^2(\Omega) \times L^2(\Omega)$ when $\varepsilon \to 0$:

$$y_\varepsilon \to y \quad \text{in } Y \text{ weak}, \quad C_0(\Omega) \text{ strong},$$

$$u_\varepsilon \to u \quad \text{in } L^2(\Omega) \text{ weak},$$

$$w_\varepsilon \to w \quad \text{in } L^2(\Omega) \text{ weak},$$

with (y, u, w) satisfying (3.6). As K and B are closed and convex in $L^2(\Omega)$ and Z we have $u \in K$ and $Ly \in B$. The form of J_ε implies that $\|w_\varepsilon + \phi(y_\varepsilon)\|_{L^2(\Omega)} \to 0$ and $\|Ty_\varepsilon - a\| \to 0$; hence $w + \phi(y) = 0$. With (3.6) this implies that (y, u) satisfies (2.1). As J is lower semicontinuous, we have that

$$\begin{aligned} J(\bar{y}, \bar{u}) &\geqq \limsup J_\varepsilon(y_\varepsilon, u_\varepsilon, w_\varepsilon) \\ &\geqq \limsup \{J(y_\varepsilon, u_\varepsilon) + \tfrac{1}{2}\|u_\varepsilon - \bar{u}\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|w_\varepsilon + \phi(\bar{y})\|^2_{L^2(\Omega)}\} \\ &\geqq J(y, u) + \tfrac{1}{2}\|u - \bar{u}\|^2_{L^2(\Omega)} + \tfrac{1}{2}\|w + \phi(\bar{y})\|^2_{L^2(\Omega)}. \end{aligned}$$

As (y, u) is feasible for (P), this implies that $u = \bar{u}$ and $w + \phi(\bar{y}) = 0$; hence $\phi(y) = \phi(\bar{y})$. With (2.1) this implies that $y = \bar{y}$. But the inequality above also implies $\|u_\varepsilon - \bar{u}\|_{L^2(\Omega)} \to 0$ and $\|w_\varepsilon + \phi(\bar{y})\|_{L^2(\Omega)} \to 0$; when we use (2.6), the result follows. □

We now are in position to prove Theorem 2, by passing to the limit in the optimality system of (P_ε).

Proof of Theorem 2. Let $(y_\varepsilon, u_\varepsilon, w_\varepsilon)$ denote a solution of (P_ε) and $(p_\varepsilon, \mu_\varepsilon, \lambda_\varepsilon)$ be given by Theorem 3. If $\{(p_\varepsilon, \mu_\varepsilon, \lambda_\varepsilon)\}$ is bounded we obtain the result with $\bar{\alpha} = 1$ by passing to the limit in the optimality system of (P_ε) with the help of Lemma 3. Now suppose that $\alpha_\varepsilon = 1/(\|p_\varepsilon\|_{L^2(\Omega)} + \|\mu_\varepsilon\|_{Z'} + \|\lambda_\varepsilon\|)$ converges toward zero. Multiplying by α_ε the optimality system given by Theorem 3 and defining

$$\bar{p}_\varepsilon = \alpha_\varepsilon p_\varepsilon, \quad \bar{\mu}_\varepsilon = \alpha_\varepsilon \mu_\varepsilon, \quad \bar{\lambda}_\varepsilon = \alpha_\varepsilon \lambda_\varepsilon,$$

we obtain, eliminating $(1/\varepsilon)(w_\varepsilon + \phi(y_\varepsilon))$ from the last equality of Theorem 3,

$$\begin{aligned} &A^* \bar{p}_\varepsilon + \phi'(y_\varepsilon)\bar{p}_\varepsilon = \alpha_\varepsilon |y_\varepsilon - y_d|^{\sigma-2}(y_\varepsilon - y_d) + T^* \bar{\lambda}_\varepsilon + L^* \bar{\mu}_\varepsilon \\ &\qquad\qquad - \alpha_\varepsilon \phi'(y_\varepsilon)(w_\varepsilon + \phi(\bar{y})) \quad \text{in } \Omega, \\ &\bar{p}_\varepsilon = 0 \quad \text{on } \Gamma, \\ &\langle \bar{\mu}_\varepsilon, z - Ly_\varepsilon \rangle \leqq 0 \quad \forall z \in B, \\ &\int_\Omega [\bar{p}_\varepsilon + \alpha_\varepsilon (Nu_\varepsilon + u_\varepsilon - \bar{u})](v - u_\varepsilon) \geqq 0 \quad \forall v \in K. \end{aligned} \tag{3.7}$$

As $\|\bar{p}_\varepsilon\|_{L^2(\Omega)} + \|\bar{\mu}_\varepsilon\|_{Z'} + \|\bar{\lambda}_\varepsilon\|$ is bounded, we may pass to the limit in the systems above by using Lemma 3; then we obtain (3.3)-(3.5), with $\bar{\alpha} = 0$ here. It remains to prove that $\bar{p} \neq 0$. If $\bar{p} = 0$, then $T^* \bar{\lambda} + L^* \bar{\mu} = 0$ by (3.3). However, (3.1) and the injectivity of T^* and L^* (by (2.5)) then imply that $\bar{\mu} = 0$ and $\bar{\lambda} = 0$. Since $\{\bar{\lambda}_\varepsilon\}$ is in $\mathbb{R}^m$ and

because of Lemma 1, we infer that $\liminf \|\bar{\mu}_\varepsilon\|_{Z'} = 0$ and $\|\bar{\lambda}_\varepsilon\| \to 0$; hence $\|\bar{p}_\varepsilon\|_{L^2(\Omega)} \to 1$. From (3.7) and Lemma 3 we deduce that $A^*\bar{p}_\varepsilon$ is bounded in $M(\Omega)$; hence $\{\bar{p}_\varepsilon\}$ is bounded in $W_0^{1,s}(\Omega)$ for all $s < n/(n-1)$. The compact injection from $W_0^{1,s}(\Omega)$ into $L^2(\Omega)$ (for $n \leqq 3$ and s close to $n/(n-1)$) implies that $\|\bar{p}_\varepsilon\|_{L^2(\Omega)} \to \|\bar{p}\|_{L^2(\Omega)} = 0$, which gives a contradiction. □

4. Applications. In this section we consider the three examples stated in § 2, and we derive the optimality system for each of them.

Example 1.

THEOREM 4. *Let $(\bar{y}, \bar{u}) \in Y \times K$ be a solution of* (P1). *Then there exist a real number $\bar{\alpha} \geqq 0$ and elements $\bar{\lambda} \in R^m$ and $\bar{p} \in W_0^{1,s}(\Omega)$ for all $s < n/(n-1)$ satisfying*

$$\bar{\alpha} + \|\bar{p}\|_{W_0^{1,s}(\Omega)} > 0, \tag{4.1}$$

$$\begin{aligned} &A\bar{y} + \phi(\bar{y}) = \bar{u} \quad \text{in } \Omega, \\ &\bar{y} = 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.2}$$

$$\begin{aligned} &A^*\bar{p} + \phi'(\bar{y})\bar{p} = \bar{\alpha}|\bar{y} - y_d|^{\sigma-2}(\bar{y} - y_d) + \sum_{i=1}^{m} \bar{\lambda}_i \delta_{[x_i]} \quad \text{in } \Omega, \\ &\bar{p} = 0 \quad \text{on } \Gamma, \end{aligned} \tag{4.3}$$

$$\int_\Omega (\bar{p} + \bar{\alpha} N\bar{u})(v - \bar{u})\, dx \geqq 0 \quad \forall v \in K. \tag{4.4}$$

Proof. Hypothesis (3.1) is trivially satisfied as $B = C_0(\Omega)$. Hence we may apply Theorem 2, which gives the result. □

In some cases it is possible to prove that the previous theorem is true with $\bar{\alpha} = 1$. We are going to study two situations where this is so.

THEOREM 5. *Let $a_{ij} \in C^2(\bar{\Omega})$, $1 \leqq i \leqq j \leqq n$. Then the results of Theorem* 2 *are obtained with $\bar{\alpha} = 1$ if Ω is connected and one of the two following hypotheses holds:*

(i) *There exists an open subset Ω_0 of Ω such that $K = K + \widetilde{L^2(\Omega_0)}$ ($\widetilde{L^2(\Omega_0)}$ is the extension by zero from $L^2(\Omega_0)$ to $L^2(\Omega)$).*

(ii) *$K = \{v \in L^2(\Omega)\colon v(x) \geqq 0$ a.e. $x \in \Omega\}$, and $u = 0$ is not optimal for* (P1).

Proof. (i) If $\bar{\alpha} = 0$, it follows from (4.3) that

$$\begin{aligned} &A^*\bar{p} + \phi'(\bar{y})\bar{p} = \sum_{i=1}^{m} \bar{\lambda}_i \delta_{[x_i]} \quad \text{in } \Omega, \\ &\bar{p} = 0 \quad \text{on } \Gamma. \end{aligned} \tag{4.5}$$

Now from (4.4) and the property of K, we get that $\bar{p} = 0$ in Ω_0. Taking $\Omega_1 = \Omega \backslash \{x_i\}_{i=1}^m$, we have

$$\begin{aligned} &A^*\bar{p} + \phi'(\bar{y})\bar{p} = 0 \quad \text{in } \Omega_1, \\ &\bar{p} = 0 \quad \text{in } \Omega_0 \backslash \{x_i\}_{i=1}^m. \end{aligned} \tag{4.6}$$

Then we can use the Prolongation Unicity Theorem (Saut and Scheurer [20]) and we deduce that $\bar{p} = 0$ in Ω_1, hence in Ω, which contradicts (4.1).

(ii) If $\bar{\alpha} = 0$, we deduce from (4.4) that $\bar{p} \geqq 0$ in Ω. If $\bar{p}$ is null on an open subset Ω_0 of Ω, we can do as in (i) and obtain a contradiction. Otherwise, for each open subset Ω_0 with $\bar{\Omega}_0$ included in Ω_1 we have

$$\max_{x \in \Omega_0} \bar{p}(x) > 0. \tag{4.7}$$

We remark that $\bar{p}$ satisfies

$$A^*\bar{p} + \max(0, \phi'(\bar{y}))\bar{p} \geqq 0 \quad \text{in } \Omega_1,$$
$$\bar{p} = 0 \quad \text{on } \Gamma.$$

Applying the Harnack inequality to $A^* + \max(0, \phi'(\bar{y}))$ (Stampacchia [21]) as in [5], we deduce that $\bar{p}(x) > 0$ everywhere in Ω_1, which with (4.4) implies that $\bar{u} = 0$ almost everywhere. □

Example 2.

THEOREM 6. *If* $(\bar{y}, \bar{u}) \in Y \times K$ *is solution of* (P2), *then there exists a real number* $\bar{\alpha} \geqq 0$ *and elements* $\bar{\mu} \in L^\infty(\Omega)$ *and* $\bar{p} \in W_0^{1,s}(\Omega)$ *such that for all* $s < n/(n-1)$:

$$\bar{\alpha} + \|\bar{p}\|_{W_0^{1,s}(\Omega)} > 0, \tag{4.8}$$

$$\begin{aligned} A\bar{y} + \phi(\bar{y}) &= \bar{u} \quad \textit{in } \Omega, \\ \bar{y} &= 0 \quad \textit{on } \Gamma, \end{aligned} \tag{4.9}$$

$$\begin{aligned} A^*\bar{p} + \phi'(\bar{y})(\bar{p}) &= \bar{\alpha}|\bar{y} - y_d|^{\sigma-2}(\bar{y} - y_d) + \bar{\mu} \quad \textit{in } \Omega, \\ \bar{p} &= 0 \quad \textit{on } \Gamma, \end{aligned} \tag{4.10}$$

$$\int_\Omega \bar{\mu}(z - \bar{y})\, dx \leqq 0 \quad \forall z \in B, \tag{4.11}$$

$$\int_\Omega (\bar{p} + \bar{\alpha} N\bar{u})(v - \bar{u})\, dx \geqq 0 \quad \forall v \in K. \tag{4.12}$$

Proof. Here again, (3.1) is satisfied because $T = 0$. Hence we may apply Theorem 2 and remark that $Z' = L^\infty(\Omega)$ and L^* is the canonical injection into $M(\Omega)$. Moreover, the regularity of $\bar{p}$ follows from (2.6), (4.10), and the fact that $\bar{\alpha}|\bar{y} - y_d|^{\sigma-2}(\bar{y} - y_d) + \bar{\mu} - \phi'(\bar{y})\bar{p}$ belongs to $L^1(\Omega)$. □

Example 3.

THEOREM 7. *If* $(\bar{y}, \bar{u}) \in Y \times K$ *is solution of* (P3), *then there exist a real number* $\bar{\alpha} \geqq 0$ *and elements* $\bar{p} \in W_0^{1,s}(\Omega)$ *for all* $s < n/(n-1)$, $\bar{\lambda} \in \mathbb{R}$, *and* $\bar{\mu} \in M(\Omega)$ *such that*

$$\bar{\alpha} + \|\bar{p}\| W_0^{1,s}(\Omega) > 0, \tag{4.13}$$

$$\begin{aligned} A\bar{y} + \phi(\bar{y}) &= \bar{u} \quad \textit{in } \Omega, \\ \bar{y} &= 0 \quad \textit{on } \Gamma, \end{aligned} \tag{4.14}$$

$$\begin{aligned} A^*\bar{p} + \phi'(\bar{y})\bar{p} &= \bar{\alpha}|\bar{y} - y_d|^{\sigma-2}(\bar{y} - y_d) + \bar{\lambda} + \bar{\mu} \quad \textit{in } \Omega, \\ \bar{p} &= 0 \quad \textit{on } \Gamma, \end{aligned} \tag{4.15}$$

$$\int_\Omega (z - \bar{y})\, d\bar{\mu} \leqq 0 \quad \forall z \in B, \tag{4.16}$$

$$\int (\bar{p} + \bar{\alpha} N\bar{u})(v - \bar{u})\, dx \geqq 0 \quad \forall v \in K. \tag{4.17}$$

Proof. We have to verify that (3.1) is satisfied. Remember that in this case L is the identity in $C_0(\Omega)$ and $T \in C_0(\Omega)'$. Take $\mu \in \partial I_B(\bar{y})$ and $\lambda \in R$ such that

$$\langle \mu, z \rangle = \langle T^*\lambda, z \rangle = \lambda \int_\Omega z\, dx \quad \forall z \in C_0(\Omega);$$

this implies that $\mu = \lambda m$, where m is the Lebesgue measure. If $\lambda \neq 0$, this implies that $\bar{y}(x) = \pm\delta$ almost everywhere, which contradicts the boundary condition. □

REFERENCES

[1] A. R. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Publishing House of Romanian Academy, Bucharest, 1978.

[3] J. F. BONNANS, *Analysis and control of a non-linear parabolic unstable system*, J. Large Scale Systems, 6 (1984), pp. 249-262.

[4] J. F. BONNANS AND E. CASAS, *Contrôle de systèmes non linéaires comportant des contraintes distribuées sur l'état*, Rapport de Recherche 300, Institut National de Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France, 1984.

[5] ———, *Contrôle de systèmes elliptiques semilinéaires comportant des contraintes distribuées sur l'état*, in Nonlinear Partial Differential Equations and Their Applications: Collège de France Seminar vol. VIII, H. Brézis and J. L. Lions, eds., Pitman, Boston, to appear, pp. 69-86.

[6] ———, *Quelques méthodes pour le contrôle optimal de problèmes comportant des contraintes sur l'état*, Anal. Stiinficice Univ. "Al. I. Cuza" din Iasi 32, Iasi, Roman: a, Matematica, 1986, pp. 58-62.

[7] ———, *On the choice of the function spaces for some state-constrained control problems*, Numer. Funct. Anal. Optim. (1984/1985), pp. 333-348.

[8] E. CASAS, *Quelques problèmes de contrôle avec contraintes sur l'état*, C.R. Acad. Sci. Paris. Sér. I, 296 (1983), pp. 509-512.

[9] ———, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309-1318.

[10] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[11] M. G. CRANDALL AND P. RABINOWITZ, *Bifurcation perturbation of simple eigenvalues, and linearized stability*, Arch. Rational Mech. Anal., 53 (1973), pp. 161-180.

[12] I. EKELAND AND R. TEMAM, *Analyse convexe et problèmes variationnels*, Dunod, Paris, 1974.

[13] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[14] V. KOMORNIK, *On the control of strongly nonlinear systems* I, Studia Sci. Math. Hungar., to appear.

[15] J. L. LIONS, *Contrôle de systèmes distribués singuliers*, Dunod, Paris, 1983.

[16] U. MACKENROTH, *Convex parabolic boundary control problems with pointwise state constraints*, J. Math. Anal. Appl., 87 (1982), pp. 256-277.

[17] ———, *On some elliptic optimal control problems with state constraints*, Optimization, 17 (1986), pp. 595-607.

[18] J. NEČAS, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[19] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

[20] J. C. SAUT AND B. SCHEURER, *Sur l'unicité du problème de Cauchy et le prolongement unique pour des équations elliptiques à coefficients non localement bornés*, J. Differential Equations, 43 (1982), pp. 28-43.

[21] G. STAMPACCHIA, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189-258.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 457-475, May 1989

ASPECTS OF POSITIVITY IN CONTROL THEORY*

TILMAN SCHANBACHER†

Abstract. This paper studies finite- and infinite-dimensional linear control systems of the form $df/dt = Af + Bu$, where A is the infinitesimal generator of a C_0-semigroup that preserves a cone C, and where Bu takes values in C. Since the reachable states are all in C, the system is not controllable in the usual sense. Of concern is "positive controllability," which means that the entire cone C can be (approximately) reached. It turns out that positive controllability is rather difficult to achieve but that for stable systems an important subclass of states can be reached. Different examples are provided.

Key words. controllability, positive controllability, positive systems, positive controls, stationary pairs

AMS(MOS) subject classifications. primary 93B05; secondary 93C05

1. Introduction. In this paper we are concerned with a linear distributed parameter control system of the form

$$\frac{d}{dt} f(t) = Af(t) + Bu(t), \qquad f(0) = f_0.$$

General questions such as controllability and stabilizability of this system have been intensively investigated in recent years by means of functional analysis, especially by the theory of strongly continuous semigroups (see, e.g., Balakrishnan [1], Curtain and Pritchard [4], [5]). Saperstone and Yorke [10], Brammer [3], Son [16], Korobov and Son [7], and Schanbacher [12] consider controllability with the additional requirement that the controls are restricted to a certain subset of the original space, thereby making particular use of the theorems of Krein and Rutman [8] on positive operators.

Until now, scarce attention has been paid to the important case where the control of a system is realizable in only one direction. In the papers on controllability with restrictions on the controls mentioned above, this problem has been taken into account, whereas the aim has still been to (approximately) reach the entire linear state space or at least a neighborhood of the initial state. However, this is possible only if the system is oscillating in some sense; in particular, it is impossible for the large class of systems described by positive semigroups. Our concern is to ask which positive states can be reached for positivity-preserving systems if the controls are taken to be positive. In analogy to the usual definition of controllability we call the system positive controllable if all positive states can be reached.

In § 2 we present the definitions and notation needed later, mainly for readers who are not familiar with the abstract theory of Banach lattices and positive operators.

Section 3 is a heuristic introduction to the problems examined in this paper. We study two examples illustrating the questions to be considered in §§ 4 and 5.

In § 4 we discuss the concept of positive controllability for positive systems, both on finite-dimensional and infinite-dimensional spaces. We obtain a neat characterization of positive controllability for finite-dimensional spaces, and show that positive controllability on infinite-dimensional spaces is rather difficult to achieve, contrary to the usual notion of controllability.

* Received by the editors October 13, 1986; accepted for publication (in revised form) May 12, 1988. This paper is part of the author's Ph.D. dissertation of the same title.

† Mathematisches Institut der Universität Tübingen, Morgenstelle 10, 7400 Tübingen, Federal Republic of Germany.

In § 5 we consider a subclass of positive states that can be reached more easily than arbitrary positive states. Existence of such states is mainly connected with asymptotic stability of the system.

We conclude §§ 4 and 5 with practical examples.

2. Preliminaries. In this section we introduce some basic notation and definitions from functional analysis following Schaefer [11] and Nagel [9].

2.1. Sets and spaces. Throughout this paper we use E and U to denote Banach spaces over the same real or complex field, endowed with a norm $\| \ \|$.

By $L(E)$, $L(U, E)$ we denote the space of linear continuous (or bounded) operators from E into E, respectively, from U into E, endowed with the canonical operator norm.

By $L^p([0, t]; U)$ $(1 \leqq p \leqq \infty)$ we denote the space of all p-integrable functions on $[0, t]$ with values in U, and by $L^p_{\mathrm{loc}}(\mathbb{R}_+; U) := \{f : \mathbb{R}_+ \to U : f_{|[0,t]} \in L^p([0, t]; U)$ for all $t > 0\}$ the space of all locally p-integrable functions on $\mathbb{R}_+$ with values in U.

Let Ω be a subset of E or U. We use the following notation:

cl Ω	the closure of Ω,
co Ω	the smallest convex set containing Ω,
cone Ω	the smallest cone containing Ω and 0,
cocone Ω	the smallest convex cone containing Ω and 0.

Following [11] we define a (real) Banach lattice E as a Banach space over $\mathbb{R}$ endowed with an order written as $\leqq$ such that $(E, \leqq)$ is a lattice and the ordering is compatible with the Banach space structure of E. We will elaborate on this.

The axioms of compatibility between the linear structure of E and the order are as follows:

$$f \leqq g \text{ implies } f + h \leqq g + h \quad \text{for all } f, g, h \text{ in } E,$$
$$f \geqq 0 \text{ implies } \lambda f \geqq 0 \quad \text{for all } f \text{ in } E \text{ and } \lambda \geqq 0.$$

Any real vector space with an ordering satisfying these two axioms is called an *ordered vector space*. The axioms imply that the set $E_+ := \{f \in E : f \geqq 0\}$ is a convex set and a cone with vertex zero, the *positive cone* of E. It follows that $f \leqq g$ if and only if $g - f \in E_+$. The elements $f \in E_+$ are called *positive*, and we write $f > 0$ if f is positive and different from zero.

An ordered vector space E is called a *vector lattice* if any two elements f, g in E have a supremum and an infimum denoted by $\sup(f, g)$, respectively, $\inf(f, g)$. For an element f of a vector lattice we write

$$|f| = \sup(f, -f) \text{ and call it the } \textit{absolute value} \text{ of } f.$$

We call two elements f, g of a vector lattice *orthogonal*, if $\inf(|f|, |g|) = 0$.

The axiom of compatibility between norm and order required for a Banach lattice is shown below:

$$|f| \leqq |g| \quad \text{implies } \|f\| \leqq \|g\|. \tag{2.1}$$

A norm on a vector lattice satisfying this axiom is called a *lattice norm*. Finally, a (real) *Banach lattice* is a Banach space E over $\mathbb{R}$ endowed with an ordering $\leqq$ such that $(E, \leqq)$ is a vector lattice and the norm on E is a lattice norm.

In particular, in Banach lattices the following formulas are valid:

$$|f + g| \leqq |f| + |g| \quad \text{and} \quad \| |f| \| = \|f\|.$$

A linear subspace I of E is called an *ideal* if $f \in I$, $|g| \leqq |f|$ implies $g \in I$.

Two ideals I, J of a vector lattice E are called *orthogonal* if any two elements $f \in I$, $g \in J$ are orthogonal. It is immediate that two ideals are orthogonal if and only if they have trivial intersection $\{0\}$. An ideal I is called *proper* if I is different from E and from $\{0\}$.

Typical ideals are all sets of the form

$$J_{f'} := \{f \in E : \langle f', |f| \rangle = 0\} \quad \text{for some } 0 \leqq f' \in E'.$$

A linear form $f' \in E'$ is called

Positive $(f' \geqq 0)$ if $\langle f', f \rangle \geqq 0$ for all $f \geqq 0$;

Strictly positive $(f' \gg 0)$ if $\langle f', f \rangle > 0$ for all $f > 0$.

A *complex Banach lattice* E is the complexification of a real Banach lattice $E_{\mathbb{R}}$ in the sense that

$$E = E_{\mathbb{R}} \oplus iE_{\mathbb{R}},$$

i.e., $E = E_{\mathbb{R}} \times iE_{\mathbb{R}}$ with scalar multiplication $(\alpha + i\beta)(f, g) = (\alpha f - \beta g, \beta f + \alpha g)$. $E_{\mathbb{R}}$ is called the *real part* of E. The absolute value of an element $h = f + ig \in E$ is defined by

$$|h| = \sup \{\cos \theta \cdot f + \sin \theta \cdot g : 0 \leqq \theta \leqq 2\pi\} \in E_{\mathbb{R}},$$

and the norm still satisfies (2.1).

Since the absolute value exists for all $f \in E$ the definition of an ideal can be extended unchanged to the complex situation. An element $f \in E$ is called positive if $f = |f|$, which means that f is a positive element of $E_{\mathbb{R}}$.

In the following, a Banach lattice denotes either a real or a complex Banach lattice.

Relevant examples of Banach lattices are given by the following spaces: $\mathbb{R}^n$, $n \in \mathbb{N}$ with the coordinatewise ordering; the function spaces $C_0(Y)$ (Y locally compact) and $L^p(\Omega)$ (Ω a measure space, $1 \leqq p \leqq \infty$) with the ordering $f \leqq g$ for $f, g \in C_0(Y)$, respectively, $L^p(\Omega)$, if $f(x) \leqq g(x)$ for all $x \in Y$, respectively, for almost all $x \in \Omega$.

2.2. Operators and semigroups. Let E and U be Banach spaces, $T \in L(E)$ and $S \in L(U, E)$. We denote the kernel of S by $\ker S$, the spectrum of T by $\sigma(T)$, the point spectrum by $p\sigma(T)$, and the resolvent set by $\rho(T)$. For all $\lambda \in \rho(T)$ we define the resolvent $R(\lambda, T) := (\lambda \operatorname{Id} - T)^{-1}$, where Id denotes the identity on E. S is called *positive* $(S \geqq 0)$ if E and U are Banach lattices and if $Su \geqq 0$ for $0 \leqq u \in U$.

Now we consider a *strongly continuous semigroup* $\mathcal{T} = (T(t))_{t \geqq 0}$ of linear continuous operators on a Banach space E, i.e.,

$$T(t+s) = T(t) \cdot T(s) \quad \text{for } t, s \geqq 0,$$

$$T(0) = \operatorname{Id},$$

$$\lim_{t \to 0} T(t)f = f \quad \text{for all } f \in E.$$

In the following we simply call $\mathcal{T}$ a *semigroup* on E.

Let $(A, D(A))$ be its *generator*, i.e.,

$$D(A) := \left\{ f \in E : \lim_{t \to 0} \frac{T(t)f - f}{t} \text{ exists} \right\},$$

$$Af := \lim_{t \to 0} \frac{T(t)f - f}{t}.$$

Then $(A, D(A))$ is a closed operator on E that determines $\mathcal{T}$ uniquely. In general, A is unbounded. We call $\omega(A) := \omega(\mathcal{T}) := \inf\{w \in \mathbb{R}$: There exists M_w such that $\|T(t)\| \leqq M_w e^{wt}$ for all $t \geqq 0\}$ the *growth bound* of A (or of $\mathcal{T}$). Then $-\infty \leqq \omega(A) < +\infty$ by [9, A-I, § 1.2].

If $A \in L(E)$ then $T(t) = \exp(tA) := \sum_{n=0}^{\infty} ((tA)^n/n!)$ by [9, A-I, 2.1].

The semigroup $\mathcal{T}$ is called *bounded* if there exists $M > 0$ such that $\|T(t)\| \leqq M$ for every $t \geqq 0$ (this implies $\omega(A) \leqq 0$).

A special class of bounded semigroups are the following: $\mathcal{T}$ is called *weakly stable* if $\langle f', T(t)f\rangle \to 0$ as $t \to \infty$ for every $f \in E$ and $f' \in E'$; *strongly stable* if $\|T(t)f\| \to 0$ as $t \to \infty$ for every $f \in E$; and *uniformly exponentially stable* if $\omega(A) < 0$. On finite-dimensional spaces these three notions of stability coincide (and we simply call the semigroup *stable*), whereas in the case of infinite-dimensional spaces these notions are strictly different ([9, A-IV, § 1]). Obviously uniform exponential stability implies strong stability, and strong stability implies weak stability. Finally we call the semigroup $\mathcal{T}$ *positive* ($\mathcal{T} \geqq 0$) if E is a Banach lattice and if $T(t)$ is a positive operator for every $t \geqq 0$.

3. Examples and questions. As a heuristic motivation for this paper we present two practical examples of linear control systems providing both the finite- and the infinite-dimensional case. In both cases we will notice that the solution of the system remains in the positive cone of a Banach lattice and we can pose two questions that will be discussed in the §§ 4 and 5, respectively. We will return to these two examples in each of these sections.

3.1. Electrically heated oven. We consider a simple model of an electrically heated oven (see [2, Ex. 1.2]) consisting of a jacket, an inner part, and a coil that directly heats the jacket and indirectly heats the interior part by means of radiation of the jacket. Let τ_0 denote the outside temperature, $\tau_1(t)$ the temperature of the jacket at time t, and $\tau_2(t)$ the temperature of the interior part at time t.

If we set

$$f(t) = \begin{pmatrix} \tau_1(t) - \tau_0 \\ \tau_2(t) - \tau_0 \end{pmatrix},$$

then the system is described by

$$\frac{d}{dt} f(t) = \begin{pmatrix} -\alpha & \beta \\ \gamma & -\delta \end{pmatrix} f(t) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} u(t), \qquad f(0) = f_0, \tag{3.1}$$

where $u(t)$ measures the heat input at time t and is always nonnegative (since it is proportional to the square of the voltage in the coil), $\alpha, \beta, \gamma, \delta$ are nonnegative constants, and f_0 is assumed to be nonnegative ($f_0 \in \mathbb{R}^2_+$). This means that the differences $\tau_1(t) - \tau_0$ and $\tau_2(t) - \tau_0$ at time $t = 0$ are nonnegative. Then it seems plausible that $f(t)$ is nonnegative for all times. Indeed, if we denote by A the matrix $\left(\begin{smallmatrix} -\alpha & \beta \\ \gamma & -\delta \end{smallmatrix}\right)$ and by B the matrix $\left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$ then by [5, 2.23] for every $u \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ a mild solution of (3.1) is given by

$$f(t) = T(t)f_0 + \int_0^t T(t-s)Bu(s)\, ds, \tag{3.2}$$

where $\mathcal{T} = (T(t))_{t \geqq 0}$ is the strongly continuous semigroup generated by A and is positive by [9, B-II, Ex. 1.4.b]. From (3.2) we see that $f(t) \in \mathbb{R}^2_+$ for all $t \geqq 0$.

The *first question* is now whether each $f_1 \in \mathbb{R}^2_+$ can be approximately reached from the origin ($f_0 = 0$) by means of a suitable nonnegative control u.

It seems possible that by means of a high input we can reach, at least for a short period of time, an arbitrarily high jacket temperature while simultaneously the interior temperature is low. However, the opposite situation, given by a low jacket and arbitrarily high interior temperature, seems to be impossible. Therefore we expect a negative answer to this question.

For our second question we consider a special class of states f_1, namely those belonging to a positive stationary pair (f_1, u_1). This means that f_1 is a constant positive solution of (3.1) for constant nonnegative input u_1 and initial state $f_0 = f_1$, i.e., $Af_1 + Bu_1 = 0$, $0 \neq f_1 \in \mathbb{R}_+^2$, $u_1 \in \mathbb{R}_+$. If we can reach these states f_1, then we can keep them constant by further input u_1. If the system is stable and if we can reach them only approximately, then it should also be possible to keep them approximately constant by further input u_1.

Therefore our *second question* follows: Can we (approximately) reach each state f_1 belonging to a positive stationary pair starting from any $f_0 \in \mathbb{R}_+^2$? The question makes sense only if there exists a positive stationary pair, and moreover if stability of the system is desirable as just mentioned.

In the concrete example and if $\gamma + \delta \neq 0$ we can show by a simple calculation the following two equivalences:

A positive stationary pair exists if and only if A has only eigenvalues with nonpositive real part.

A positive stationary pair (f_1, u_1) with $u_1 \neq 0$ exists if and only if A has only eigenvalues with negative real part (i.e., A is stable).

In practice (see [2, Ex. 1.2]) A is stable and the parameters α, β, γ, δ are (strictly) positive.

3.2. Heat equation on a finite rod with noninsulated ends. We consider the heat equation on a rod of length 1 with noninsulated ends given by

$$\frac{\partial}{\partial t}\chi(t, x) = \frac{\partial^2}{\partial x^2}\chi(t, x), \quad 0 \leqq x \leqq 1, \quad t \geqq 0,$$

$$\chi(t, 0) = \chi(t, 1) = 0,$$

$$\chi(0, x) = \chi_0(x),$$

where χ_0 is a given nonnegative function on $[0, 1]$.

Again we wish to control the system by a nonnegative one-dimensional input, i.e., by a function $\psi(x) \cdot u(t)$, where $\psi(x) \geqq 0$ for $0 \leqq x \leqq 1$, $u(t) \geqq 0$ for $t \geqq 0$. As in the first example, we can interpret this control as an electrical heating input that for all time is proportional to a given heat distribution ψ. Then we obtain

$$\begin{aligned} &\frac{\partial}{\partial t}\chi(t, x) = \frac{\partial^2}{\partial x^2}\chi(t, x) + \psi(x) \cdot u(t), \quad 0 \leqq x \leqq 1, \quad t \geqq 0, \\ &\chi(t, 0) = \chi(t, 1) = 0, \\ &\chi(0, x) = \chi_0(x), \end{aligned} \tag{3.3}$$

and again it seems plausible that χ remains nonnegative for all times.

We state this problem as an abstract control problem on the space $E := L^2[0, 1]$. Assume $\psi \in E$ and let $U = \mathbb{R}$, $B \in L(U, E)$, $B\alpha := \alpha \cdot \psi$, $f_0 := \chi_0$ and let $A = d^2/dx^2$ be the operator on E with domain $D(A) = \{f \in E : f'' \in E, f(0) = f(1) = 0\}$. It is known [5, Eq. 3.9], [9, C-II, Ex. 1.5.b] that $(A, D(A))$ generates a positive semigroup

$\mathcal{T} = (T(t))_{t \geqq 0}$ on E given by

$$T(t)f = \sum_{n=1}^{\infty} e^{-n^2\pi^2 t}\langle f, e_n\rangle e_n,$$

where $f \in E$, $e_n(x) = \sqrt{2} \cdot \sin(n\pi x)$ for $0 \leqq x \leqq 1$, $\langle f, e_n\rangle := \int_0^1 f(x)e_n(x)\,dx$. Then (3.3) can be written as (set $f(t) := \chi(t, \cdot) \in E$):

$$\frac{d}{dt} f(t) = Af(t) + Bu(t),$$

$$f(t) \in D(A),$$

$$f(0) = f_0$$

and for $u \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ a mild solution is again given by (3.2).

We observe that the semigroup $\mathcal{T}$ is uniformly exponentially stable:

$$\|T(t)f\| \leqq \left(\sum_{n=1}^{\infty} e^{-2n^2\pi^2 t}\langle f, e_n\rangle^2\right)^{1/2}$$

$$\leqq e^{-\pi^2 t}\left(\sum_{n=1}^{\infty} \langle f, e_n\rangle^2\right)^{1/2} = e^{-\pi^2 t}\|f\|$$

and we claim that a positive stationary pair exists.

Proof. Let

$$f(x) := -\int_0^x \int_0^y \psi(z)\,dz\,dy + x\int_0^1 \int_0^y \psi(z)\,dz\,dy,$$

where ψ is chosen as above; then for $f \in D(A)$ we obtain $Af(x) = -\psi(x)$. Assume that f is not contained in the positive cone of E. Since $f(0) = f(1) = 0$ implies that f has a negative minimum in some $x_0 \in (0, 1)$ we obtain

$$0 = f'(x_0) = -\int_0^{x_0} \psi(z)\,dz + \int_0^1 \int_0^y \psi(z)\,dz\,dy,$$

and since $\psi \geqq 0$

$$f'(x) = -\int_0^x \psi(z)\,dz + \int_0^1 \int_0^y \psi(z)\,dz\,dy \geqq f'(x_0) = 0 \quad \text{for } 0 \leqq x \leqq x_0,$$

$$f(x_0) = f(0) + \int_0^{x_0} f'(x)\,dx = 0 + \int_0^{x_0} f'(x)\,dx \geqq 0,$$

which is a contradiction. Hence $f \geqq 0$, $Af + B1 = 0$, which means that $(f, 1)$ is a positive stationary pair, □

Therefore it might be interesting to consider the same two questions as in the first example.

4. Positive controllability. The notion of controllability usually is defined in the sense that we want to reach a dense subset of the entire state space or at least of a neighborhood of zero. However, in many instances for systems with restrictions on the controls, it is a priori known that all reachable states are contained in a closed cone C of the state space (as in the examples of the previous section). In this case controllability in the former sense is impossible but it is interesting to know conditions under which the reachable states are dense in C (which was precisely our first question in § 3). In the present section we characterize systems for which the reachable states are contained in the positive cone of a Banach lattice, and then try to find conditions

for a particular class under which a dense subset of the positive cone can be reached. This is defined as positive controllability; we will see that this implies the usual controllability. We will be able to give feasible criteria for positive controllability for finite-dimensional state spaces and strong necessary conditions for infinite-dimensional state spaces and finite-dimensional input spaces. We conclude this section with some applications.

Let E and U be Banach spaces, let $(A, D(A))$ be the generator of a strongly continuous semigroup $\mathcal{T} = (T(t))_{t \geq 0}$ on E, and $B \in L(U, E)$. We consider the following linear control system:

$$\frac{d}{dt} f(t) = Af(t) + Bu(t), \qquad f(0) = f_0, \tag{4.1}$$

where $f(t) \in E$, $f_0 \in E$, $u(t) \in U$.

By [5, 2.23] for every $u \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ a mild solution of (4.1) is given by

$$f(t) = T(t)f_0 + \int_0^t T(t-s)Bu(s)\, ds. \tag{4.2}$$

We recall the following definition.

DEFINITION 4.1. Let $1 \leqq p \leqq \infty$ and consider (4.1). We call E the *state space* and U the *input space*. Define for $t > 0$ the set of *reachable* states from the origin ($f_0 = 0$) in time t as

$$R_t := \left\{ \int_0^{t'} T(t'-s)Bu(s)\, ds \colon u \in L^p([0, t']; U), 0 \leqq t' \leqq t \right\},$$

and define the set of reachable states from the origin in arbitrary time as $R = \bigcup_{t>0} R_t$. The pair (A, B) is called

Exactly controllable in time t if $R_t = E$;

Exactly controllable if $R = E$;

Approximately controllable in time t if R_t is dense in E;

Approximately controllable if R is dense in E.

The following proposition for finite-dimensional state spaces is known as the "rank condition."

PROPOSITION 4.2 [4, 11.4]. *If* $\dim E = n < \infty$, $\dim U = m < \infty$, *then all controllability notions in* (4.1) *are equivalent to the fact that the* $n \times nm$*-matrix* $(B, AB, A^2B, \cdots, A^{n-1}B)$ *has rank* n.

Now we assume that E is a Banach lattice, that u in (4.1) takes values in a subset Ω of U and that f_0 is an element of the positive cone E_+ of E. First we want to know conditions under which the solution $f(t)$ of (4.1) remains in E_+ for all such u and f_0. We state the following proposition.

PROPOSITION 4.3. *Let E be a Banach lattice, let U be a Banach space, let $(A, D(A))$ be the generator of a strongly continuous semigroup $\mathcal{T} = (T(t))_{t \geq 0}$ on E, $B \in L(U, E)$, and $0 \in \Omega \subseteq U$. The following assertions are equivalent*:

(i) *The mild solutions of* (4.1) *remain in E_+ for every $f_0 \in E_+$ and every $u \in L^1_{\mathrm{loc}}(\mathbb{R}_+, U)$ with $u(t) \in \Omega$ for $t \geqq 0$.*

(ii) *$\mathcal{T}$ is a positive semigroup and $B\Omega \subseteq E_+$.*

Proof. The implication (ii) $\rightarrow$ (i) follows immediately from the variation of constants formula (4.2).

(i) $\rightarrow$ (ii): Since $0 \in \Omega$ we can take $u(\cdot)$ identical zero, then $f(t) = T(t)f_0 \in E_+$ for all $f_0 \in E_+$, i.e., $\mathcal{T}$ is positive. On the other hand, taking $f_0 = 0$, $u(s) = u/t$ for some $u \in \Omega$ and all $0 \leqq s \leqq t$, we obtain by (4.2) $f(t) = 1/t \int_0^t T(s)Bu\, ds$, which tends to Bu as $t \rightarrow 0$. The closedness of E_+ implies $Bu \in E_+$. □

After this characterization we restrict our considerations to the case where U is also a Banach lattice, Ω the positive cone U_+, and therefore B is a positive operator by the preceding proposition. Then we define positive controllability as follows.

DEFINITION 4.4. Let E and U be Banach lattices, let $(A, D(A))$ be the generator of a positive semigroup $\mathcal{T} = (T(t))_{t \geqq 0}$ on E, and take $B \in L(U, E)$ positive. For $t > 0$ and $1 \leqq p \leqq \infty$ the set of *reachable* states from the origin in time t by means of nonnegative controls u is defined as

$$R_t^+ := = \left\{ \int_0^{t'} T(t'-s)Bu(s)\, ds \colon u \in L^p([0, t']; U),\ u(s) \in U_+,\ 0 \leqq s \leqq t' \leqq t \right\},$$

and the set of reachable states from the origin in arbitrary time by means of nonnegative controls u as $R^+ := U_{t>0} R_t^+$. The pair (A, B) is called:

Exactly positive controllable in time t if $R_t^+ = E_+$;

Exactly positive controllable if $R^+ = E_+$;

Approximately positive controllable in time t if $\operatorname{cl}(R_t^+) = E_+$;

Approximately positive controllable if $\operatorname{cl}(R^+) = E_+$.

Of course, each notion implies the corresponding controllability notion in Definition 4.1 since $R_t = R_t^+ - R_t^+$. The notion of approximate positive controllability (in time t) is independent of p and we can even restrict the controls to C^∞-functions.

We give some simple examples.

Example 4.5. (a) Let $E = U$ be a Banach lattice, $A = 0$, $B = \mathrm{Id}$. Then obviously (A, B) is exactly positive controllable in any time $t > 0$.

(b) Let $E = U = L^p(\mathbb{R}_+)$, $1 \leqq p < \infty$, and $\mathcal{T}$ the translation semigroup on E given by $(T(t)f)(x) = f(x+t)$ for all x, $t \geqq 0$. Let $(A, D(A))$ be its generator and $B = \mathrm{Id}$. Then (A, B) is exactly positive controllable in any time $t > 0$.

Proof. A strongly continuous semigroup $\mathcal{S} = (S(t))_{t \geqq 0}$ on E is given by

$$S(t)f(x) = \begin{cases} f(x-t) & \text{for } x \geqq t, \\ 0 & \text{for } x < t, \end{cases}$$

and $T(t)S(t) = \mathrm{Id}$ for all $t \geqq 0$. For $f \in E_+$ we set $u(s) := (1/t)S(t-s)f$ and obtain $f = \int_0^t (1/t) f\, ds = \int_0^t T(t-s)u(s)\, ds \in R_t^+$.

Example 4.6. Let $E = \mathbb{R}^2$, $U = \mathbb{R}$, $A = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right)$. (A, B) is exactly controllable in any time $t > 0$ by Proposition 4.2. However, (A, B) is not approximately positive controllable. Since

$$\exp(tA) = \begin{pmatrix} e^t & 0 \\ 0 & e^{2t} \end{pmatrix}$$

we see that $R^+ \subseteq \left\{ \left(\begin{smallmatrix} a \\ b \end{smallmatrix}\right) \in \mathbb{R}_+^2 \colon a \leqq b \right\}$.

We state the following proposition, which gives a useful representation of the sets $\operatorname{cl}(R_t^+)$ and $\operatorname{cl}(R^+)$. For a subset M of a Banach space E let $\operatorname{co} M$, $\operatorname{cocone} M$ be defined as in § 2.

PROPOSITION 4.7. *Let the assumptions be as in Definition* 4.4.

(a) $\operatorname{cl}(R_t^+) = \operatorname{cl}(\operatorname{co}\{T(s)Bu \colon 0 \leqq s \leqq t,\ u \in U_+\})$,
$\operatorname{cl}(R^+) = \operatorname{cl}(\operatorname{co}\{T(s)Bu \colon 0 \leqq s,\ u \in U_+\})$.

(b) *Let $U = \mathbb{R}^m$ and let $e(1), \cdots, e(m)$ be the canonical unit vectors of $\mathbb{R}^m$. Then*
$\operatorname{cl}(R_t^+) = \operatorname{cl}(\operatorname{cocone}\{T(s)Be(i) \colon 0 \leqq s \leqq t,\ 1 \leqq i \leqq m\})$,
$\operatorname{cl}(R^+) = \operatorname{cl}(\operatorname{cocone}\{T(s)Be(i) \colon 0 \leqq s,\ 1 \leqq i \leqq m\})$.

Proof. (a) By definition of the integral we see that the sets on the left-hand side are contained in the sets on the right. On the other hand, R_t^+, R^+ are convex since U_+

is convex; it remains to show that $T(s)Bu \in \operatorname{cl}(R_t^+)$ for all $0 \leqq s \leqq t$, $t > 0$, $u \in U_+$. For this let $u \in U_+$ and

$$u_n(s) = \begin{cases} nu & \text{for } 0 \leqq s \leqq 1/n, \\ 0 & \text{for } 1/n \leqq s \leqq t. \end{cases}$$

Then $u_n(\cdot) \in L^\infty([0, t]; U)$ and

$$\left\| \int_0^t T(t-s)Bu_n(s)\, ds - T(t)Bu \right\| \leqq n \cdot \int_0^{1/n} \| T(t-s)Bu - T(t)Bu \|\, ds \to 0 \text{ as } n \to \infty.$$

(b) Since $B\mathbb{R}_+^m = \operatorname{cocone}\{Be(1), \cdots, Be(m)\}$ we see that (a) implies (b). □

We apply Proposition 4.7 to the following example.

Example 4.8. Let $E = U = l^p$, $1 \leqq p < \infty$, and let the operators A, B be given by

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ \vdots & & & \vdots & & & \vdots \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots \\ \vdots & & & \vdots & & & \vdots \end{bmatrix}.$$

Obviously $A \in L(E)$, $B \in L(U, E)$ and A generates the positive semigroup $\mathcal{T} = (T(t))_{t \geqq 0}$ given by

$$T(t) = \begin{bmatrix} 1 & t & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & t & 0 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & t & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & & & \vdots & & & \vdots \end{bmatrix}.$$

Let $e(1)$, $e(2), \cdots$ denote the canonical unit vectors of E and U. Then for $i \in \mathbb{N}$, $T(0)Be(i) = e(2i)$ and $\lim_{t \to \infty} (T(t)Be(i)/\| T(t)Be(i) \|) = e(2i-1)$. Hence $\operatorname{cl}(R^+) = E_+$ by Proposition 4.7(a), i.e., (A, B) is approximately positive controllable. However, for all $t > 0$ the ratio of the $2i$th coordinate of $\int_0^t T(t-s)Bu(s)\, ds$ to the $(2i-1)$th coordinate is greater than $1/t$; hence $e(2i-1)$ cannot be contained in $\operatorname{cl}(R_t^+)$ or R^+, i.e., (A, B) is neither approximately positive controllable in any finite time $t > 0$ nor exactly positive controllable. □

4.1. The finite-dimensional case. The following theorem gives a characterization for approximate positive controllability on finite-dimensional state spaces.

THEOREM 4.9. *Let the assumptions be as in Definition* 4.4, $E = \mathbb{R}^n$, $U = \mathbb{R}^m$, *and* $e[1], \cdots, e[n]$, $e(1), \cdots, e(m)$ *the canonical unit vectors of* E, *respectively,* U.

(a) (A, B) *is approximately positive controllable in time* $t > 0$ *if and only if*

(4.3) *For all* $1 \leqq i \leqq n$ *there exists* $1 \leqq j \leqq m$ *and* $\mu > 0$ *such that* $e[i] = \mu Be(j)$.

(b) (A, B) *is approximately positive controllable if and only if*

(4.4) *For all* $1 \leqq i \leqq n$ *there exist* $1 \leqq j \leqq m$ *and* $\mu > 0$ *such that* $e[i] = \mu Be(j)$ *or* $e[i] = \lim_{t \to \infty} (T(t)Be(j)/\| T(t)Be(j) \|)$.

Proof. We first remark that by [13] $\lim_{t\to\infty}(T(t)f/\|T(t)f\|)$ exists for every $0<f\in\mathbb{R}^n$.

The sufficiency of (a) and (b) then is obvious by Proposition 4.7(b).

To show necessity we may assume, as in the proof of Lemma A2 in the Appendix, that $Be(j)\neq 0$ for all $1\leqq j\leqq m$.

(a) If (A, B) is approximately positive controllable in some time $t>0$, then by Lemma A2(a)

$$\{T(s)Be(j)\colon 0\leqq s\leqq t,\ 1\leqq j\leqq m\}\cap J_{f'}\neq\varnothing \quad \text{for all } J_{f'}\neq\{0\}. \tag{4.5}$$

Since $A+\|A\|\,\mathrm{Id}\geqq 0$ [9, C-II, Thm. 1.11], we obtain for every $s>0$ and $1\leqq j\leqq m$

$$\begin{aligned} e^{s\|A\|}\cdot T(s)Be(j) &= Be(j)+\sum_{n=1}^{\infty}\frac{1}{n!}(s(A+\|A\|\,\mathrm{Id}))^n Be(j)\\ &\geqq Be(j). \end{aligned} \tag{4.6}$$

Now (4.5) and (4.6) yield $\{Be(j)\colon 1\leqq j\leqq m\}\cap J_{f'}\neq\varnothing$ for all $J_{f'}\neq\{0\}$. This implies the assertion (4.3).

(b) If (A, B) is approximately controllable, then by Lemma A2(b)

$$\left[\{T(s)Be(j)\colon 0\leqq s,\ 1\leqq j\leqq m\}\cup\left\{\lim_{t\to\infty}\frac{T(t)Be(j)}{\|T(t)Be(j)\|}\colon 1\leqq j\leqq m\right\}\right]\cap J_{f'}\neq\varnothing$$

for all $J_{f'}\neq\{0\}$. Again by (4.6) we obtain

$$\left[\{Be(j)\colon 1\leqq j\leqq m\}\cup\left\{\lim_{t\to\infty}\frac{T(t)Be(j)}{\|T(t)Be(j)\|}\colon 1\leqq j\leqq m\right\}\right]\cap J_{f'}\neq\varnothing$$

for $J_{f'}\neq\{0\}$ and this implies the assertion (4.4). □

COROLLARY 4.10. *Let the assumptions be as in the preceding theorem.*

(a) *(A, B) is approximately positive controllable in time $t>0$ if and only if $BU_+=E_+$. In particular, approximate positive controllability in time $t>0$ implies $m\geqq n$.*

(b) *If (A, B) is approximately positive controllable then $2m\geqq n$, and for every $1\leqq i\leqq n$ the following holds: $e[i]$ is an eigenvector of A or $e[i]=\mu Be(j)$ for some $1\leqq j\leqq m$, $\mu>0$.*

(c) *If (A, B) is exactly positive controllable then (A, B) is approximately positive controllable in every time $t>0$.*

Proof. (a) If $Bu=e[i]$ for $u>0$ then $\mu Be(j)=e[i]$ for some $\mu>0$, $1\leqq j\leqq m$ since B is positive. The assertion follows from Theorem 4.9(a).

(b) By [13] $\lim_{t\to\infty}(T(t)f/\|T(t)f\|)$ is for every $0<f\in\mathbb{R}^n$ an eigenvector of A; therefore the assertion follows from Theorem 4.9(b).

(c) We show that (4.3) holds. Since (A, B) is exactly positive controllable there exists for all $J_{f'}\neq\{0\}$ and $0<f\in J_{f'}$ a time $t>0$ such that $f\in R_t^+$; by Lemma A2(a), we obtain that $\{T(s)Be(j)\colon 0\leqq s\leqq t,\ 1\leqq j\leqq m\}\cap J_{f'}\neq\varnothing$. As in the proof of Theorem 4.9(a) this yields (4.3). □

We point out that Theorem 4.9 and Corollary 4.10 show that the different positive controllability notions do not coincide on finite-dimensional state spaces, contrary to the usual controllability notions (Proposition 4.2).

As an example for these results we mention again Example 4.6. By Theorem 4.9(b) it is immediate that (A, B) in (4.6) is not approximately positive controllable. More examples will be discussed below.

4.2. The infinite-dimensional case. The results of Corollary 4.10 for finite-dimensional state spaces may lead to the conjecture that positive controllability is

impossible if the state space is infinite-dimensional and the input space is finite-dimensional. In fact, we will show that approximate positive controllability in finite time $t>0$ is never possible for a large class of systems on nearly all state spaces of practical interest. Moreover, on the same state spaces, approximate positive controllability is impossible for systems for which (as in the finite-dimensional case) $\lim_{t\to\infty}(T(t)Bf/\|T(t)Bf\|)$ exists for all $f>0$. However, as pointed out in [13], the limit may not exist, in which case there exists an example where the system is approximately positive controllable.

Finally we should mention that by [5, 3.2] and by the remark in Definition 4.4 (that positive controllability implies the usual controllability) exact positive controllability is never possible in the case where the state space is infinite-dimensional and the input space finite-dimensional.

We state the following theorem.

THEOREM 4.11. *Let E be an infinite-dimensional Banach lattice with a strictly positive linear form $f'\in E'$. Let $(A, D(A))$ be the generator of a positive semigroup $\mathcal{T}=(T(t))_{t\geq 0}$ on E and take $B\in L(\mathbb{R}^m, E)$ positive, $m\in\mathbb{N}$. Denote by $e(1),\cdots,e(m)$ the canonical unit vectors of $\mathbb{R}^m$ and assume $Be(j)\neq 0$ for $1\leq j\leq m$.*

(a) *(A, B) is not approximately positive controllable in time $t>0$ if $T(s)Be(j)\neq 0$ for every $0\leq s\leq t$ and $1\leq j\leq m$.*

(b) *(A, B) is not approximately positive controllable if $T(s)Be(j)\neq 0$ for every $0\leq s$ and $1\leq j\leq m$ and if $\lim_{t\to\infty}(T(t)Be(j)/\|T(t)Be(j)\|)$ exists for every $1\leq j\leq m$.*

We will prove the theorem after Remark 4.12.

Remark 4.12. The assumption that there exists a strictly positive linear form is fulfilled for all state spaces of practical interest, e.g., $C_0(X)$, X a locally compact subset of $\mathbb{R}^n$, $L^p(\Omega, \lambda)$, Ω a measurable subset of $\mathbb{R}^n$, λ the Lebesgue measure.

The assumption $Be(j)\neq 0$ in the theorem is made only for convenience. If $Be(j)=0$ for some j we modify the conditions stated in the theorem:

(a) $0\notin\{T(s)Be(j): 0\leq s\leq t, 1\leq j\leq m, Be(j)\neq 0\}$,

(b) $0\notin\{T(s)Be(j): 0\leq s, 1\leq j\leq m, Be(j)\neq 0\}$ and $\lim_{t\to\infty}(T(t)Be(j)/\|T(t)Be(j)\|)$ exists for all $1\leq j\leq m$ for which $Be(j)\neq 0$.

For the proof of this fact we restrict B to the sublattice of $\mathbb{R}^m$ generated by those $e(j)$ for which $Be(j)\neq 0$, and then apply Theorem 4.11.

The condition in Theorem 4.11(a), as well as the first condition in Theorem 4.11(b), is obviously fulfilled by every positive semigroup that can be extended to a (not necessarily positive) strongly continuous *group* and also by every *analytic* positive semigroup. Let $\mathcal{T}=(T(t))_{t\geq 0}$ be an analytic positive semigroup and $T(t)f=0$ for some $t>0$, $f\in E$; then $T(s)f=0$ for $s\geq t$, and hence $\langle f', T(s)f\rangle=0$ for every $s\geq t$ and $f'\in E'$. Since $\mathcal{T}$ is analytic this implies that $\langle f', T(s)f\rangle=0$ for all $s>0$ and $f'\in E'$; hence $\langle f', f\rangle=0$ and $f=0$.

Proof of Theorem 4.11. (a) As mentioned in the proof of Lemma A2(a), the set $C:=\{T(s)Be(j): 0\leq s\leq t, 1\leq j\leq m\}$ is compact and does not contain zero by assumption. On the other hand, by Lemma A4 there exist infinitely many pairwise orthogonal ideals $J_{f'}\neq\{0\}$, $0<f'\in E'$, and by Lemma A.3, C can only intersect finitely many of them. By Lemma A2(a) this implies that $\mathrm{cl}(R_t^+)$ has nontrivial intersection with only finitely many of these ideals; hence $\mathrm{cl}(R_t^+)\neq E_+$, i.e., (A, B) is not approximately positive controllable in time t.

(b) Again, as mentioned in the proof of Lemma A2(b), the assumption implies that

$$\left\{\frac{T(s)Be(j)}{\|T(s)Be(j)\|}: 0\leq s, 1\leq j\leq m\right\}\cup\left\{\lim_{t\to\infty}\frac{T(t)Be(j)}{\|T(t)Be(j)\|}: 1\leq j\leq m\right\}$$

is compact and does not contain zero. Lemmas A4, A3, and A2(b) yield as in (a) that $\mathrm{cl}(R^+) \neq E_+$, i.e., (A, B) is not approximately positive controllable. □

We do not know whether we can drop the condition in Theorem 4.11(a) as well as the condition that there exists a strictly positive linear form. In other words, we do not know of any example where E is infinite-dimensional, $U = \mathbb{R}^m$, and where (A, B) is positive controllable in some finite time $t > 0$. But the following example shows that (A, B) may be approximately positive controllable if the limit in the assumption of Theorem 4.11(b) does not exist. The basic idea of the example is due to Zabczyk [17].

Example 4.13. Let $E = L^p(\mathbb{R}_+)$, $1 \leqq p < \infty$, $U = \mathbb{R}$, $\mathcal{T} = (T(t))_{t \geqq 0}$ being the (positive) translation semigroup on E given by

$$(T(t)f)(x) = f(x+t) \quad \text{for } x, t \geqq 0$$

with generator $(A, D(A))$, and let $B \in L(\mathbb{R}, E)$ be a positive operator given by

$$Bu = u \cdot b \quad \text{for some } b \in E_+.$$

We claim that (A, B) is approximately positive controllable for a suitable $b \in E_+$, which we will construct as follows.

First we know that there exists a countable subset S of normalized positive functions with compact support such that $\mathrm{cl}(\mathrm{cocone}\, S) = E_+$: Take, for example, all characteristic functions on intervals $[q, q+r]$, $q, r \in Q_+ \setminus \{0\}$ divided by their L^p-norm. In S there exists a sequence $f_1, f_2, \cdots$ in which each element of S occurs infinitely often.

By assumption the support $\mathrm{spt}(f_i)$ is contained in $[0, t_i]$ for some $t_i > 0$. Define

$$s(1) := 0, \quad s(n+1) := \sum_{i=1}^{n} t_i \quad \text{for } n \geqq 1$$

and let

$$f_i^{(s)} \text{ denote the function } t \to \begin{cases} f_i(t-s) & \text{for } t \geqq s, \\ 0 & \text{for } t < s. \end{cases}$$

Since $\mathrm{spt}(f_i^{(s(i))}) \cap \mathrm{spt}(f_j^{(s(j))}) \subseteq [s(i), s(i+1)] \cap [s(j), s(j+1)]$ has measure zero for $i \neq j$, we can define

$$b := \sum_{i=1}^{\infty} 2^{-i^2} \cdot f_i^{(s(i))}$$

and obtain $\|b\| \leqq (\sum_{i=1}^{\infty} 2^{-pi^2})^{1/p} \leqq 1$; hence $b \in E_+$.

Because of Proposition 4.7(a) (A, B) is approximately positive controllable for this b if $S \subseteq \mathrm{cl}\{T(s)bu\colon 0 \leqq s,\ u \geqq 0\}$. For this we now show that for each $f \in S$ and $n \in \mathbb{N}$ there exists $s, u > 0$ such that

$$\|T(s)bu - f\| < 2^{-n}. \tag{4.7}$$

Since f occurs infinitely often in the sequence (f_i), we can find $k > n$ such that $f = f_k$. Let $s = s(k)$, $u = 2^{k^2}$. Then by definition of $\mathcal{T}$ we obtain

$$T(s)f^{(s(i))} = T(s(i) + t_i + \cdots + t_{k-1})f^{(s(i))} = 0 \quad \text{for } i < k,$$

$$T(s)f^{(s(k))} = f,$$

$$T(s)f^{(s(i))} = T(s(k))f^{(s(k)+s(i)-s(k))} = f^{(s(i)-s(k))} \quad \text{for } i > k.$$

Hence

$$T(s)bu - f = \sum_{i=k+1}^{\infty} 2^{k^2-i^2} \cdot f^{(s(i)-s(k))}.$$

Since

$$\operatorname{spt}(f_i^{(s(i)-s(k))}) \cap \operatorname{spt}(f_j^{(s(j)-s(k))}) \subseteq [s(i)-s(k), s(i+1)-s(k)] \cap [s(j)-s(k), s(j+1)-s(k)]$$

has measure zero for $i \neq j$, we obtain

$$\|T(s)bu-f\| \leqq \left(\sum_{i=k+1}^{\infty} 2^{p(k^2-i^2)}\right)^{1/p} = \left(\sum_{i=k+1}^{\infty} 2^{p(k+i)(k-i)}\right)^{1/p} = \left(\sum_{i=1}^{\infty} 2^{-p(2k+i)i}\right)^{1/p}$$
$$\leqq \left(\sum_{i=1}^{\infty} 2^{-p2k} \cdot 2^{-pi^2}\right)^{1/p} \leqq 2^{-2k}\left(\sum_{i=1}^{\infty} 2^{-pi^2}\right)^{1/p} \leqq 2^{-2k} < 2^{-n}.$$

Thus (4.7) is proved. □

4.3. Applications. We apply the main Theorems 4.9 and 4.11 to some concrete examples.

Example 4.14. The electrically heated oven of § 3 is neither approximately positive controllable in any time $t>0$ nor exactly positive controllable because of Corollary 4.10(a), (c). It is approximately positive controllable by Theorem 4.9(b) if and only if

$$\binom{0}{1} = \lim_{t\to\infty}\left(T(t)\binom{1}{0} \Big/ \left\|T(t)\binom{1}{0}\right\|\right). \tag{4.8}$$

By [13] a necessary condition for (4.8) is that $\binom{0}{1}$ is an eigenvector to A, i.e., $\beta = 0$. We verify that

$$T(t) = \begin{pmatrix} e^{-\alpha t} & 0 \\ \gamma t\, e^{-\alpha t} & e^{-\delta t} \end{pmatrix} \quad \text{if } \alpha = \delta,$$

$$T(t) = \begin{pmatrix} e^{-\alpha t} & 0 \\ \gamma(e^{-\alpha t} - e^{-\delta t}) \cdot (\delta - \alpha)^{-1} & e^{-\delta t} \end{pmatrix} \quad \text{if } \alpha \neq \delta.$$

Hence (4.8) holds if and only if

$$\gamma \neq 0, \quad \beta = 0, \quad \delta \leqq \alpha.$$

In the physical situation we have $\alpha, \beta, \gamma, \delta > 0$ and therefore the oven is not approximately positive controllable. But by Proposition 4.2, (A, B) is (exactly) controllable if and only if the matrix $\binom{1\ -\alpha}{0\ \ \gamma}$ is nonsingular, i.e., $\gamma \neq 0$.

Example 4.15. In [6] the following system is discussed as an approximation for a distributed network:

$$\frac{d}{dt}f(t) = A_0 f(t) + \mathrm{Id}\,[\psi(f_1(t)), \cdots, \psi(f_n(t))], \tag{4.9}$$

where

$$A_0 = \begin{bmatrix} -2 & 1 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ 0 & 0 & 1 & -2 & \cdots & 0 \\ & & & \cdots & & \\ 0 & 0 & & \cdots & -2 & 1 \\ 0 & 0 & & \cdots & 1 & -2 \end{bmatrix},$$

$f(t)=(f_1(t),\cdots,f_n(t))\in\mathbb{R}^n=:E$, and ψ is a real function on $\mathbb{R}$. If ψ is linear, then (4.9) reads as follows:

$$\frac{d}{dt}f(t)=A_0f(t)+\mu\ \mathrm{Id}\,f(t) \tag{4.10}$$

for some $\mu\in\mathbb{R}$.

Define $A:=A_0+\mu$ Id. Then by [9, B-II, Ex. 1.4.b], A generates a positive semigroup and we can consider the linear control system (4.1) for some positive $B\in L(\mathbb{R}^m,\mathbb{R}^n)$, $m\in\mathbb{N}$. Since none of the canonical unit vectors $e[i]$ of $\mathbb{R}^n$ is an eigenvector of A, we infer from [13] that the second condition in Theorem 4.9(b) can never be fulfilled. Hence (A, B) is approximately positive controllable if and only if it is approximately positive controllable for any time $t>0$, i.e., if and only if $BU_+=E_+$ (by Corollary 4.10(a)). However, from Proposition 4.2 it is immediate that (A, B) is (exactly) controllable if BU contains, e.g., $e[1]$ or $e[n]$, in particular, B may even have rank 1.

Example 4.16. We consider the heat equation on a finite rod with Dirichlet boundary conditions as in § 3. As shown in [14], $\lim_{t\to\infty}(T(t)f/\|T(t)f\|)$ exists for every $f>0$; in particular, $T(t)f\neq 0$ for all $t\geqq 0$, $f>0$. Since there exists a strictly positive linear form on $L^2[0, 1]$, by Theorem 4.11 we obtain that the control system is not approximately positive controllable.

In the same way we can show, using the results of [13] and [14], that the system is also not approximately positive controllable if we replace the Dirichlet by Neumann boundary conditions.

Example 4.17. Consider the Laplacian on $\mathbb{R}^n$. This means we consider the control problem (4.1) for $E=C_0(\mathbb{R}^n)$ or $E=L^p(\mathbb{R}^n)$, $1\leqq p<\infty$, $U=\mathbb{R}^m$, $0<B\in L(U, E)$, $D(A)=\{f\in E:\Delta f\in E\}$, $Af=\Delta f$. By [9, C-II, Ex. 1.5.c and d and Remark] $(A, D(A))$ generates a positive contraction semigroup $\mathcal{T}=(T(t))_{t\geqq 0}$ on E given by

$$T(t)f(x)=(4\pi t)^{-n/2}\int_{\mathbb{R}^n}\exp(-|x-y|^2/4t)\circ f(y)\,dy,$$

where $|x-y|$ denotes the Euclidean norm of $x-y$ in $\mathbb{R}^n$.

This shows that $T(t)f(x)>0$ for all $x\in\mathbb{R}^n$ whenever $f>0$ and the assumptions of Theorem 4.11(a) are fulfilled. Hence (A, B) is not approximately positive controllable in any time $t>0$. On the other hand, (A, B) may be approximately controllable in every time $t>0$ even for $U=\mathbb{R}^2$ (see [5, 3.17]).

5. Positive stationary pairs. In § 4 we have obtained somewhat negative answers to the first question posed in § 3. However, it is often not so important to reach the entire positive cone of the state space. It suffices to design a nonnegative control by which particular positive states can be approximated and held constant for all times. This was just the problem formulated in the second question of § 3 for positive states belonging to positive stationary pairs. Moreover, we have observed in § 3 a relationship between existence of a positive stationary pair and stability of the system. In this section we will discuss these problems systematically.

Let E and U be Banach lattices, let $\mathcal{T}=(T(t))_{t\geqq 0}$ be a positive semigroup on E with generator $(A, D(A))$, and take $B\in L(U, E)$ positive. As in the previous section we consider the control problem (4.1), where $u\in L^1_{\mathrm{loc}}(\mathbb{R}_+, U)$ takes values in U_+ and $f_0\in E_+$.

We call a pair $(f_1, u_1)\in(E_+\setminus\{0\})\times U_+$ *positive stationary* if $Af_1+Bu_1=0$. In this case $f(\cdot)\equiv f_1$ is a nonzero constant solution of (4.1) for $u(\cdot)\equiv u_1$, $f_0=f_1$. We are interested in whether each $f_1\in E_+$ belonging to a positive stationary pair can be

(approximately) reached and held constant by a nonnegative control $u(\cdot)$. Moreover, we study the connection between existence of positive stationary pairs and stability of the semigroup $\mathcal{T}$ as defined in § 2.

Our main results will be formulated in Theorem 5.1 and 5.6, and applications will be given at the end of the section.

We state the following theorem.

THEOREM 5.1. *Let E and U be Banach lattices, let $\mathcal{T}=(T(t))_{t\geq 0}$ be a uniformly exponentially stable positive semigroup on E with generator $(A, D(A))$, and take $B\in L(U, E)$ positive. Then we have the following*:

(a) *To each $u_1\in U_+\backslash\ker B$ there exists exactly one $f_1\in E_+$ such that (f_1, u_1) is a positive stationary pair.*

(b) *If (f_1, u_1) is a positive stationary pair, $f_0\in E_+$ and $u(\cdot)\equiv u_1$, then the solution of (4.1) tends to f_1 as $t\to\infty$.*

Proof. (a) If $\mathcal{T}$ is uniformly exponentially stable, then $0\in\rho(A)$ and $-A^{-1}$ is positive by [9, C-III, Thm. 1.1]. For all $u_1\in U_+\backslash\ker B$ we therefore obtain that $(-A^{-1}Bu_1, u_1)$ is a positive stationary pair. On the other hand, $0\in\rho(A)$ implies that for each $u_1\in U_+$, there exists at most one f_1 such that (f_1, u_1) is a positive stationary pair.

(b) For a positive stationary pair (f_1, u_1) and $u(\cdot)\equiv u_1$, we obtain for the mild solution of (4.1) that $f(t)=T(t)f_0-\int_0^t T(t-s)Af_1\,ds$. This is equal to $T(t)(f_0-f_1)+f_1$, and therefore tends to f_1 as $t\to\infty$ since $\mathcal{T}$ is stable. □

From the proof above we obtain the following corollary.

COROLLARY 5.2. *Let E and U be Banach lattices, let $\mathcal{T}=(T(t))_{t\geq 0}$ be a weakly stable positive semigroup on E with generator $(A, D(A))$, and take $B\in L(U, E)$ positive. Then we have the following*:

(a) *For each $u_1\in U_+\backslash\ker B$ there exists at most one $f_1\in E_+$ such that (f_1, u_1) is a positive stationary pair; such $f_1\in E_+$ exists if $0\in\rho(A)$.*

(b) *If (f_1, u_1) is a positive stationary pair $f_0\in E_+$ and $u(\cdot)\equiv u_1$, then the solution of (4.1) tends to f_1 in the weak topology of E as $t\to\infty$; the convergence holds in the norm-topology if $\mathcal{T}$ is strongly stable.*

Proof. (a) The uniqueness of f_1 follows from the fact that A is injective if $\mathcal{T}$ is weakly stable. Moreover, weak stability implies that the growth bound $\omega(A)$ of A is not positive and therefore, by [9, C-III, Thm. 1.1], the resolvent $R(\lambda, A)$ of A in λ is a positive operator for all $\lambda>0$. Thus if $0\in\rho(A)$ we have $-A^{-1}=\lim_{\lambda\to 0}R(\lambda, A)\geqq 0$, and as in the proof of Theorem 5.1(a) we obtain the existence of a positive stationary pair.

(b) This follows immediately from the proof of Theorem 5.1(b). □

For another immediate consequence of Theorem 5.1 we recall the definition of R^+ from Definition 4.4.

COROLLARY 5.3. *Let E and U be Banach lattices, let $\mathcal{T}=(T(t))_{t\geq 0}$ be a positive semigroup on E with generator $(A, D(A))$, and take $B\in L(U, E)$ positive. Then $\{R(\lambda, A)Bu\colon u\in U_+, \lambda>\omega(A)\}\subseteq \mathrm{cl}\,(R^+)$.*

Proof. Obviously R^+ does not change if we multiply $T(t)$ by $e^{-\lambda t}$ for $\lambda>\omega(A)$. Then $(e^{-\lambda t}T(t))_{t\geq 0}$ is an exponentially stable semigroup with generator $A_\lambda:=A-\lambda$, and $-(A_\lambda)^{-1}=R(\lambda, A)$. The assertion then follows from the proof of Theorem 5.1. □

The following example shows that there may not exist a positive stationary pair if the semigroup $\mathcal{T}$ is strongly stable only.

Example 5.4. We consider the Laplacian on $E=C_0(\mathbb{R})$ or $L^p(\mathbb{R})$, $1\leqq p<\infty$ as in Example 4.17 with $B\in L(\mathbb{R}^m, E)$ positive. The semigroup $\mathcal{T}$ generated by $(A, D(A))$ is strongly stable [9, A-IV, Ex. 1.2]. We claim that there does not exist a positive

stationary pair. Suppose that $Af+Bu=0$ for some $0\leqq f\in D(A)$, $u\geqq 0$. We show $f=0$.

The equation $Af+Bu=0$ implies that the second derivative f'' of f takes only nonpositive values. Now for some constants $c_1, c_2\in\mathbb{R}$ the function f is given by

$$f(x)=c_2+c_1x+\int_0^x\int_0^y f''(z)\,dz\,dy, \qquad x\in\mathbb{R}. \tag{5.1}$$

Let g be the function that maps y on $\int_0^y f''(z)\,dz$. Then $g(0)=0$ and g is antitone. Since $g(y)\leqq 0$ for $y\geqq 0$ and since $f\geqq 0$, we obtain by (5.1) (if we choose x to be large) that $c_1\geqq 0$. On the other hand, $g(y)\geqq 0$ for $y\leqq 0$ implies that $\int_0^x\int_0^y f''(z)\,dz\,dy=-\int_x^0 g(y)\,dy\leqq 0$ for $x\leqq 0$. Equation (5.1) and $f\geqq 0$ imply $c_1\leqq 0$; hence $c_1=0$.

Since g is antitone, $\alpha:=\lim_{y\to-\infty} g(y)$ and $\beta:=\lim_{y\to+\infty} g(y)$ exist in the extended real field and $\alpha\geqq 0$, $\beta\leqq 0$. Assume $\alpha>0$; then $\int_0^x\int_0^y f''(z)\,dz\,dy=-\int_x^0 g(y)\,dy\to-\infty$ as $x\to-\infty$, which contradicts the facts that $f\geqq 0$ and $c_1=0$. In the same way the case $\beta<0$ is impossible. Thus $\alpha=\beta=0$, i.e., $g=0$; hence $f(x)=c_2$ for all x. Since $f\in C_0(\mathbb{R})$ or $f\in L^p(\mathbb{R})$, this is only possible if $f=0$. □

Thus we have shown that strong stability is not sufficient to ensure the existence of a positive stationary pair. On the other hand, the following example shows that not even boundedness of the semigroup is necessary for the existence of such a pair.

Example 5.5. Let $E=\mathbb{R}^2$, $U=\mathbb{R}$, $A=\left(\begin{smallmatrix}-1 & 0\\ 0 & 1\end{smallmatrix}\right)$, $B=\left(\begin{smallmatrix}1\\ 0\end{smallmatrix}\right)$. Then $\left(\left(\begin{smallmatrix}1\\ 0\end{smallmatrix}\right), 1\right)$ is a positive stationary pair, but $(e^{tA})_{t\geqq 0}$ is an unbounded semigroup since $1\in p\sigma(A)$.

The following theorem states the conditions under which the existence of a positive stationary pair implies boundedness or even stability of the semigroup $\mathcal{T}$.

THEOREM 5.6. *Let E and U be Banach lattices, and let $\mathcal{T}=(T(t))_{t\geqq 0}$ be a positive semigroup on E with generator $(A, D(A))$. Take $B\in L(U, E)$ positive and let (f_1, u_1) be a positive stationary pair.*

(a) *$\mathcal{T}$ is bounded if $f_1\in \operatorname{int} E_+$.*

(b) *$\mathcal{T}$ is stable if E is finite-dimensional, $f_1\in \operatorname{int} E_+$, and $Bu_1\in \operatorname{int} E_+$.*

Proof. (a) If f_1 is an interior point of E_+, then to each $f\in E_+$ there exists $\lambda_f>0$ such that $f\leqq \lambda_f\cdot f_1$. Since $f_1-T(t)f_1=-\int_0^t T(s)Af_1\,ds=\int_0^t T(s)Bu_1\,ds\geqq 0$ we obtain $0\leqq T(t)f\leqq \lambda_f T(t)f_1\leqq \lambda_f f_1$ and $\|T(t)f\|\leqq \lambda_f\|f_1\|$ for all $t\geqq 0$. By the uniform boundedness principle T is bounded.

(b) Since $\|T(t)f\|\leqq \lambda_f\|T(t)f_1\|$ for every $f\in E_+$, it suffices to show that $T(t)f_1\to 0$ as $t\to\infty$. Now $T(t)f_1=f_1-\int_0^t T(s)Bu_1\,ds$; hence

$$T(r+t)f_1=T(r)f_1-\int_r^{t+r} T(s)Bu_1\,ds\leqq T(r)f_1 \quad \text{for } r\geqq 0.$$

Thus $(T(t)f_1)_{t\geqq 0}$ is decreasing and therefore converges to some $g\in E_+$ as $t\to\infty$ since E is finite-dimensional [11, II, Thm. 5.11]. Then g is a fixed vector of every $T(t)$ and $0\leqq g\leqq T(t)f_1$ for all $t\geqq 0$. Moreover $Bu_1\in \operatorname{int} E_+$ implies $1/t\int_0^t T(s)Bu_1\,ds\in \operatorname{int} E_+$ for small $t>0$. Since

$$\frac{1}{t}(f_1-g)\geqq\frac{1}{t}(f_1-T(t)f_1)=\frac{1}{t}\int_0^t T(s)Bu_1\,ds,$$

we obtain $1/t(f_1-g)\in \operatorname{int} E_+$. Thus, $f_1-g\in \operatorname{int} E_+$; hence there exists $\mu>0$ such that $f_1\leqq \mu(f_1-g)$, and therefore

$$0\leqq T(t)f_1\leqq \mu T(t)(f_1-g)=\mu(T(t)f_1-g)\to 0 \quad \text{as } t\to\infty.$$

This implies $T(t)f_1\to 0$ as $t\to\infty$. □

We remark that Theorem 5.6 can be applied only to spaces E that are isomorphic to $C(X)$, X compact, since for all other spaces $\operatorname{int} E_+=\varnothing$ [11, II, § 7].

As we have seen in Example 5.5, $\mathcal{T}$ may not be bounded if (f_1, u_1) is a positive stationary pair with $f_1 \notin \operatorname{int} E_+$. Moreover, $\mathcal{T}$ may not be strongly stable if $f_1 \in \operatorname{int} E_+$ and $Bu_1 \notin \operatorname{int} E_+$, as the following example shows.

Example 5.7. Let $E = \mathbb{R}^2$, $U = \mathbb{R}$, $A = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$; then $\left(\left(\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right), 1\right)$ is a positive stationary pair but $(\exp(tA))_{t \geqq 0}$ is not stable since $0 \in p\sigma(A)$.

We apply our results to the examples of § 3.

Example 5.8. We consider the electrically heated oven from § 3. As mentioned there, A is stable if and only if there exists a positive stationary pair (f_1, u_1) with $u_1 > 0$.

Since $\alpha, \beta, \gamma, \delta \geqq 0$ it is immediate that A is stable if and only if $\det A > 0$. Then all positive stationary pairs are scalar multiples of $\left((\det A)^{-1}\left(\begin{smallmatrix} \delta \\ \gamma \end{smallmatrix}\right), 1\right)$.

Moreover, in § 3 we have seen that for $\gamma + \delta \neq 0$ a positive stationary pair $(x_1, 0)$ exists if and only if A is not stable but has eigenvalues with nonpositive real parts, i.e., $\det A = 0$. Then all positive stationary pairs are scalar multiples of $\left(\left(\begin{smallmatrix} \delta \\ \gamma \end{smallmatrix}\right), 0\right)$.

In the physical situation we have $\det A > 0$ (cf. 4.14) and Theorem 5.1 is applicable.

Example 5.9. We consider the heat equation of § 3. As shown there the semigroup is uniformly exponentially stable and a positive stationary pair exists (cf. Theorem 5.1(a)). All f_1 belonging to a positive stationary pair (f_1, u_1) can be approximately reached by the control $u(\cdot) \equiv u_1$ according to Theorem 5.1(b).

Appendix A. The following four lemmas on order ideals are needed in § 4. We recall the definition of the ideals $J_{f'}$ from § 2.

LEMMA A1. *Let C be a compact subset of the positive cone E_+ of a Banach lattice E, and let $0 < f' \in E'$ be a positive linear form. If there exists $0 < f \in \operatorname{cl}(\operatorname{cocone} C) \cap J_{f'}$, then $C \cap J_{f'} \neq \emptyset$.*

Proof. The assertion is trivial if $0 \in C$. Therefore let $0 \notin C$ and

$$f = \lim_{m \to \infty} \sum_{i=1}^{n(m)} \alpha_{i,m} \cdot f_{i,m} \quad \text{for some } \alpha_{i,m} > 0, \quad f_{i,m} \in C, \quad n(m) \in \mathbb{N}.$$

We set

$$\beta_{i,m} := \alpha_{i,m} \|f_{i,m}\| > 0, \qquad g_{i,m} := f_{i,m} \cdot \|f_{i,m}\|^{-1}$$

and obtain

$$0 = \langle f', f \rangle = \lim_{m \to \infty} \sum_{i=1}^{n(m)} \alpha_{i,m} \langle f', f_{i,m} \rangle$$

$$= \lim_{m \to \infty} \sum_{i=1}^{n(m)} \beta_{i,m} \langle f', g_{i,m} \rangle.$$

Moreover,

$$\sum_{i=1}^{n(m)} \beta_{i,m} = \sum_{i=1}^{n(m)} \beta_{i,m} \|g_{i,m}\| \geqq \left\| \sum_{i=1}^{n(m)} \beta_{i,m} \cdot g_{i,m} \right\| \to \|f\| \quad \text{as } t \to \infty.$$

Thus we can assume that

$$\sum_{i=1}^{n(m)} \beta_{i,m} \langle f', g_{i,m} \rangle < \frac{\|f\|}{2m} \quad \text{and} \quad \sum_{i=1}^{n(m)} \beta_{i,m} \geqq \frac{\|f\|}{2} \quad \text{for every } m \in \mathbb{N}.$$

Hence for every m there exists $i(m)$ such that $\langle f', g_{i(m),m} \rangle < 1/m$. Since C is compact we can find a subsequence of $(f_{i(m),m})_m$ that converges to some $f_0 \in C$. Hence a

subsequence $(h_m)_m$ of $(g_{i(m),m})_m$ converges to $g := f_0 \cdot \|f_0\|^{-1}$, and $\langle f', g\rangle = \lim_{m\to\infty}\langle f', h_m\rangle = 0$. Thus we obtain $\langle f', f_0\rangle = 0$, $f_0 \in J_{f'}$. □

LEMMA A2. *Let the assumptions be as in Definition* 4.4, *let* $U = \mathbb{R}^m$, $e(1), \cdots, e(m)$ *be the canonical unit vectors of* $\mathbb{R}^m$, *and* $0 < f' \in E'$. *Then the following assertions are valid.*

(a) *If* $0 < f \in \mathrm{cl}\,(R_t^+) \cap J_{f'}$ *then*

$$\{T(s)Be(i)\colon 0 \leqq s \leqq t,\ 1 \leqq i \leqq m,\ Be(i) \neq 0\} \cap J_{f'} \neq \varnothing.$$

(b) *Assume that* $0 \notin \{T(s)Be(i)\colon 0 \leqq s,\ 1 \leqq i \leqq m,\ Be(i) \neq 0\}$ *and that* $\lim_{t\to\infty}(T(t)Be(i)/\|T(t)Be(i)\|)$ *exists for every* $1 \leqq i \leqq m$ *for which* $Be(i) \neq 0$. *If* $0 < f \in \mathrm{cl}\,(R^+) \cap J_{f'}$ *then*

$$\left[\{T(s)Be(i)\colon 0 \leqq s,\ 1 \leqq i \leqq m,\ Be(i) \neq 0\} \cup \left\{\lim_{t\to\infty}\frac{T(t)Be(i)}{\|T(t)Be(i)\|}\colon 1 \leqq i \leqq m,\ Be(i) \neq 0\right\}\right] \cap J_{f'} \neq \varnothing.$$

Proof. We can assume that $Be(i) \neq 0$ for all $1 \leqq i \leqq m$; otherwise we restrict B to the sublattice of $\mathbb{R}^m$ spanned by those vectors $e(i)$ for which $Be(i) \neq 0$. This will not affect the set R_t^+ and the assumptions of the lemma remain valid.

(a) $\{T(s)Be(i)\colon 0 \leqq s \leqq t\}$ is compact for $1 \leqq i \leqq m$ since $\mathcal{T}$ is strongly continuous. Set $C := \{T(s)Be(i)\colon 0 \leqq s \leqq t,\ 1 \leqq i \leqq m\}$ and apply Proposition 4.7(b) and Lemma A1 to obtain the assertion.

(b) By Proposition 4.7(b) we know that the set

$$\mathrm{cocone}\left[\left\{\frac{T(s)Be(i)}{\|T(s)Be(i)\|}\colon 0 \leqq s,\ 1 \leqq i \leqq m\right\} \cup \left\{\lim_{t\to\infty}\frac{T(t)Be(i)}{\|T(t)Be(i)\|}\colon 1 \leqq i \leqq m\right\}\right]$$

is dense in R^+. Now continuity of $t \to (T(t)Be(i)/\|T(t)Be(i)\|)$ and the existence of $\lim_{t\to\infty}(T(t)Be(i)/\|T(t)Be(i)\|)$ imply compactness of the set

$$\left\{\frac{T(s)Be(i)}{\|T(s)Be(i)\|}\colon 0 \leqq s,\ 1 \leqq i \leqq m\right\} \cup \left\{\lim_{t\to\infty}\frac{T(t)Be(i)}{\|T(t)Be(i)\|}\colon 1 \leqq i \leqq m\right\}.$$

By Lemma A1 we know that this set has nonvoid intersection with $J_{f'}$. The assertion follows immediately since $T(t)Be(i)/\|T(t)Be(i)\| \in J_{f'}$ if and only if $T(t)Be(i) \in J_{f'}$. □

LEMMA A3. *Let* E *be a Banach lattice, let* C *be a compact subset of* E_+, $0 \notin C$, *and let* $(I_i)_{i\in J}$ *be a class of proper ideals of* E *that are pairwise orthogonal. Then* $C \cap I_i = \varnothing$ *for almost all* $i \in J$ (*i.e., for all except finitely many*).

Proof. Assume that C has nontrivial intersection with infinitely many ideals I_i. Since $0 \notin C$ there exists a sequence of pairwise orthogonal elements f_n with $\inf_{n\in\mathbb{N}}(\|f_n\|) > \alpha > 0$. But this implies by [11, II, Prop. 1.4] that $\|f_n - f_m\| > \alpha$ for all $n \neq m$ in contradiction to the compactness of C. □

LEMMA A4. *Let* E *be an infinite-dimensional Banach lattice. Then the following assertions are equivalent*:

(i) *There exists an infinite number of pairwise orthogonal ideals of the form* $J_{f'} := \{f \in E\colon \langle f', |f|\rangle = 0\}$, $J_{f'} \neq \{0\}$, $0 < f' \in E'$.

(ii) *There exist two orthogonal ideals of the form* $J_{f'}$, $J_{f'} \neq \{0\}$, $0 < f' \in E'$.

(iii) *There exists a strictly positive linear form* f' *on* E.

The proof of the lemma is due to Räbiger [18] and can be found in [15]. □

Acknowledgments. I thank Professors R. Nagel and J. Zabczyk for their stimulation and guidance in the preparation of this paper.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Second edition, Springer-Verlag, Berlin, New York, 1981.
[2] S. BARNETT, *Introduction to Mathematical Control Theory*, Clarendon Press, Oxford, 1975.
[3] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, SIAM J. Control, 10 (1972), pp. 339-353.
[4] R. F. CURTAIN AND A. J. PRITCHARD, *Functional Analysis in Modern Applied Mathematics*, Academic Press, London, New York, 1977.
[5] ———, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, Berlin, New York, 1978.
[6] P. GRABOWSKI, *Contribution to Stability Theory of Nonlinear Distributed Networks*, preprint, 1986.
[7] V. I. KOROBOV AND N. K. SON, *Controllability of linear systems in Banach spaces with restrictions on the control* II, Differential'nye Uravneniye, 16 (1980), pp. 1010-1022. (In Russian.) Differential Equations, 16 (1980), pp. 633-642. (In English.)
[8] M. G. KREIN AND M. A. RUTMAN, *Linear operators leaving invariant a cone in a Banach space*, Uspekhi Mat. Nauk (N.S.) 3 (1948), pp. 3-95; Translations Series 1, Vol. 10, pp. 199-326, American Mathematical Society, Providence, RI, 1962.
[9] R. NAGEL, *One-Parameter Semigroups of Positive Operators*, Springer-Verlag, Berlin, New York, 1986.
[10] S. H. SAPERSTONE AND J. A. YORKE, *Controllability of linear oscillatory systems with positive controls*, SIAM J. Control, 9 (1971), pp. 253-262.
[11] H. H. SCHAEFER, *Banach Lattices and Positive Operators*, Springer-Verlag, Berlin, New York, 1974.
[12] T. SCHANBACHER, *Starkstetige Halbgruppen in der Kontrolltheorie*, Semesterbericht Funktionalanalysis, Tübingen, Winter 1983/84, pp. 73-82.
[13] ———, *Asymptotic behaviour of positive semigroups*, Math. Z., 195 (1987), pp. 481-485.
[14] ———, *Asymptotic behaviour of positive semigroups*, Semesterbericht Funktionalanalysis, Tübingen, Summer 1986, pp. 91-104.
[15] ———, *Aspects of positivity in control theory*, Ph.D. thesis, University of Tübingen, 1986.
[16] N. K. SON, *Local controllability of linear systems with restrained controls in Banach space*, Acta Math. Vietnam, 5 (1980), pp. 78-87.
[17] J. ZABCZYK, personal communication.
[18] F. RÄBIGER, personal communication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 476-494, May 1989

002

NEW RESULTS IN THE REDUCTION OF LINEAR TIME-VARYING DYNAMICAL SYSTEMS*

J. ZHU†‡ AND C. D. JOHNSON†

Abstract. This paper considers finite-dimensional, linear time-varying dynamical systems (LDS) of the form $\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}$, $\mathbf{x}(t_0) = \mathbf{x}_0$. Such systems are said to be *semiproper* (self-commuting) if for all t, τ, $\mathbf{A}(t)\mathbf{A}(\tau) = \mathbf{A}(\tau)\mathbf{A}(t)$. Using some recent results for obtaining explicit solutions for semiproper systems (see [26]-[29], [34], [35]), the family of Lyapunov reducible systems is expanded to include those systems that can be reduced to semiproper ones via what will be called D-similarity transformations. Within this new framework is defined the notion of primary D-similarity transformations, and every LDS that is "well defined" in a certain sense is proved reducible by a finite sequence of primary D-similarity transformations. The paper also presents an explicit technique for constructing such transformations for LDS with virtually triangular $\mathbf{A}(t)$ (i.e., $\mathbf{A}(t) = \mathbf{L}\mathbf{T}(t)\mathbf{L}^{-1}$ for some nonsingular constant matrix $\mathbf{L}$ and triangular matrix $\mathbf{T}(t)$). There are difficulties in obtaining, explicitly, primary D-similarity transformations for the reduction of general LDS. However, this paper shows that, instead of studying such general cases, it suffices to investigate only the reduction of LDS with *normal* $\mathbf{A}(t)$. To achieve these results, $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}\} = \mathbf{A}\mathbf{x} - \dot{\mathbf{x}}$ is treated as an operator on a vector space over a differential field, and thereby some familiar results in the theory of matrices are generalized over a number field. In particular, the authors introduce the notions of partial spaces, partially linear operators, linear differential equation (LDE) operators $\mathscr{P}_{\mathbf{A}}$, and D-similarity transformations, all of which are believed to be new.

Key words. linear dynamical system, time-varying system, Lyapunov reduction, semiproper reduction, differential algebra, matrices over a differential field

AMS(MOS) subject classifications. primary 34A05; secondary 93B17, 93B28, 93B40, 34A30, 93C05, 15A33, 15A57

1. Introduction and overview of the main results.

This paper concerns the class of finite-dimensional, homogeneous linear differential equations, with variable coefficient matrix, having the form

$$\dot{\mathbf{x}} = \mathbf{A}(t)\mathbf{x}, \qquad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{1.1}$$

This type of differential equation is the mathematical model for finite-dimensional linear time-varying dynamical systems when the entries $a_{ij}(t)$ in $\mathbf{A}(t)$ are real functions of a real variable t in time. We shall denote a fundamental solution matrix of (1.1) by $\mathbf{X}_{\mathbf{A}}(t)$, and denote the state-transition matrix of (1.1) by $\boldsymbol{\phi}_{\mathbf{A}}(t, t_0)$, which is given by

$$\boldsymbol{\phi}_{\mathbf{A}}(t, t_0) = \mathbf{X}_{\mathbf{A}}(t)\mathbf{X}_{\mathbf{A}}^{-1}(t_0). \tag{1.2}$$

Because of well-known difficulties in obtaining analytical solutions and stability information for systems (1.1), it is desirable to transform (1.1), by some invertible transformation, to a simpler form that can be solved analytically or studied qualitatively by existing techniques. At the end of the last century, Lyapunov introduced the important notion of a reducible system [14] which can be stated as follows.

* Received by the editors August 31, 1987; accepted for publication (in revised form) May 19, 1988. The first author's research was supported by the Li Foundation, New York prior to 1984, and by the Department of Electrical and Computer Engineering, University of Alabama, Huntsville, Alabama, since 1984.

† Department of Electrical and Computer Engineering, University of Alabama, Huntsville, Alabama 35899.

‡ Department of Automatic Control, Beijing Polytechnic University, Beijing, People's Republic of China.

DEFINITION 1.1 (Lyapunov). A linear system (1.1) is said to be reducible, in the sense of Lyapunov, if there exist real numbers M, N, and a nonsingular matrix $\mathbf{L}(t)=[l_{ij}(t)]$ satisfying:

(i) $|l_{ij}(t)| \leqq M < \infty$ and $|dl_{ij}(t)/dt| \leqq M < \infty$ for all i, j and for all $t > t_0$,

(ii) $|\det \mathbf{L}(t)| > N > 0$, for all $t > t_0$

such that the matrix

$$\mathbf{B} = \mathbf{L}^{-1}(t)\mathbf{A}(t)\mathbf{L}(t) - \mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t) \tag{1.3}$$

is constant.

The transformation (1.3) with $\mathbf{L}(t)$ satisfying (i) and (ii) is known as a Lyapunov transformation. It is well known that all periodic systems (1.1) are reducible in the sense of Lyapunov and, moreover, that the stability of such systems can be studied qualitatively via Floquet theory [5]. However, general criteria for the reducibility of nonperiodic systems (1.1) are not known. Moreover, even if a system (1.1) is known to be reducible, it is not clear how to find a Lyapunov transformation to accomplish that reduction.

In this paper, we introduce a new and somewhat broader point of view for investigating reducibility and generating reduction transformations for systems (1.1). It is recalled that a matrix function $\mathbf{A}(t)$ is said to be semiproper (self-commutative) on an interval $I \subseteq \mathbb{R}$ if $\mathbf{A}(t)\mathbf{A}(\tau) = \mathbf{A}(\tau)\mathbf{A}(t)$, for all $t, \tau \in I$. A linear system (1.1) is said to be semiproper if $\mathbf{A}(t)$ is semiproper. Note that, in particular, all time-varying systems (1.1) with *diagonal* or *Jordan* coefficient matrices, and all systems (1.1) with *constant* coefficient matrices, are semiproper. Recently, we have developed some new results for deriving finite-form analytical solutions $\boldsymbol{\phi}_{\mathbf{A}}(t, t_0)$ and stability criteria [26], [27], [34], [35] for the class of semiproper systems (1.1). Therefore it is natural to extend the family of Lyapunov reducible systems to include those systems that can be reduced to semiproper ones via Lyapunov transformations. In this regard, it should be noted that the classical boundedness requirements (i) and (ii) for Lyapunov transformations were traditionally imposed to preserve stability properties of the original system. However, those classical requirements are often overly restrictive and may rule out many candidate equivalence transformations that would enable derivation of explicit solutions and/or stability information for (1.1). Therefore it is desirable to relax requirements (i) and (ii) in Definition 1.1, to the extent possible.

Motivated by the foregoing argument, we introduce the following (new) definition of a "semiproper reducible" system.

DEFINITION 1.2. A linear system (1.1) is said to be *semiproper reducible* if there exists a matrix $\mathbf{L}(t)$ satisfying

$$\det \mathbf{L}(t) \equiv \text{constant} \neq 0 \tag{1.4}$$

such that the matrix

$$\mathbf{B}(t) = \mathbf{L}^{-1}(t)\mathbf{A}(t)\mathbf{L}(t) - \mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t) \tag{1.5}$$

is semiproper.

Remarks. (1) In this paper it will be shown in Theorem 5.2 that every Lyapunov reducible system (1.1) is semiproper reducible.

(2) Note that (1.4) assures the existence of $\mathbf{L}^{-1}(t)$ as well as the *same* divergence rates for $\mathbf{L}(t)$ and $\mathbf{L}^{-1}(t)$. The intrinsic significance of imposing condition (1.4) will be further explored in the sequel, and in separate papers [30], [31].

Hereafter, we refer to the transformation defined by (1.4) and (1.5) as a D-similarity transformation. A linear system (1.1) will be called *well defined* if $\mathbf{A}(t)$ is locally Lebesgue integrable.

In this paper, we employ Definition 1.2 to prove the following important theorem regarding semiproper reducibility.

THEOREM 1.1. *For every well-defined linear system* (1.1) *there exists a diagonal matrix* $\mathbf{D}(t)$, *and a matrix* $\mathbf{L}(t)$ *having nonzero constant determinant, such that*

$$\mathbf{D}(t)=\mathbf{L}^{-1}(t)\mathbf{A}(t)\mathbf{L}(t)-\mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t). \tag{1.6}$$

To find explicitly the D-similarity transformation matrix $\mathbf{L}(t)$ in (1.6), we will introduce in § 4 the notions of left and right D-elementary operations and D-elementary matrices as a natural generalization of the elementary operations and elementary matrices in the theory of matrices over a number field. By this means we will prove the following important theorem characterizing D-similarity transformations.

THEOREM 1.2. *Every matrix* $\mathbf{L}(t)$ *having nonzero constant determinant can be written as a product of a finite number of D-elementary matrices.*

Theorem 1.2 suggests that every D-elementary matrix constitutes a D-similarity transformation. Such a D-similarity transformation will be called *primary.* Using this terminology and the results of Theorems 1.1 and 1.2, we will establish the following fundamental result for semiproper reducible systems.

THEOREM 1.3. *Every well-defined linear system* (1.1) *is reducible to a* (*diagonal*) *semiproper system by a finite sequence of primary D-similarity transformations.*

Although Theorem 1.3 is an exciting theoretical result, its practical application is hampered because procedures for finding the primary D-similarity transformations involve, in general, solving systems of Riccati-type nonlinear differential equations that are usually as difficult to solve as the original ones. However, it is not necessary to study such general cases because we shall prove, as another of the main results in this paper, that every well-defined linear system (1.1) can be "decomposed" into two subsystems of form (1.1) with *normal* coefficient matrices. Moreover, the solution of the original system can be expressed in terms of the solutions of those two normal subsystems. It suffices, therefore, to investigate only semiproper reductions for normal systems.

Despite the difficulties mentioned above, for those systems (1.1) with virtually triangular $\mathbf{A}(t)$ (i.e., $\mathbf{A}(t)=\mathbf{L}\mathbf{T}(t)\mathbf{L}^{-1}$ for some nonsingular constant matrix $\mathbf{L}$ and triangular matrix $\mathbf{T}(t)$), the primary D-similarity transformations can always be found explicitly. The procedures for doing so, along with some illustrative examples, are developed in this paper.

To achieve our results with mathematical rigor, we shall first take the matrix functions $\mathbf{A}(t)$ as operators on a vector space over a differential field, and generalize some of the familiar results in the theory of matrices over a number field (note that a number field can be viewed as a special case of differential fields with the usual derivative operation). Those results are then relaxed so they can be applied to well-defined matrices $\mathbf{A}(t)$. As a byproduct of this approach, we obtain some additional results, which are interesting and important in their own right.

2. Vector spaces over a differential field. In this section, we establish the mathematical foundation for our main results. It is assumed that the reader is familiar with the algebraic notions such as fields, vector spaces over a number field, linear operators on a vector space, etc. Using those concepts, we adopt the notion of a *differential field* [11] defined as follows.

DEFINITION 2.1 [11]. A field $\mathbb{F}$ with addition and multiplication $\{+,\cdot\}$ is called an *ordinary differential field* if $\mathbb{F}$ adopts a derivation operator $\boldsymbol{\delta}$, and $\mathbb{F}$ is closed under

the derivative operation defined by

(a) $\boldsymbol{\delta}\{x+y\}=\boldsymbol{\delta}\{x\}+\boldsymbol{\delta}\{y\}$,

(b) $\boldsymbol{\delta}\{x\cdot y\}=\boldsymbol{\delta}\{x\}\cdot y+x\cdot\boldsymbol{\delta}\{y\}$

for all $x, y\in\mathbb{F}$. An element $C\in\mathbb{F}$ is said to be *constant* if $\boldsymbol{\delta}\{C\}=0$. All constants in $\mathbb{F}$ form a field called the *constant subfield* $\mathbb{F}_c$ of $\mathbb{F}$.

Without causing confusion, in the sequel we will simply call $\mathbb{F}$ a differential field, or D-field. Note that, with the ordinary derivative operation, the number fields $\mathbb{R}, \mathbb{C}$ are (trivial) D-fields. Nontrivial examples of D-field are the field of rational functions and the field of meromorphic functions in a complex domain.

Because of the lack of a multiplicative inverse for the zero element in the constant subfield $\mathbb{F}_c$, a D-field $\mathbb{F}$ cannot contain functions vanishing on a subinterval of positive measure. Hampered by this constraint, many functions having important applications, such as piecewise constant functions ($\mathbb{C}^\infty$ functions), cannot be studied using the powerful notion of the D-field. For this reason we now introduce the notion of an *augmented differential field* defined as follows.

DEFINITION 2.2. The union of a field $\mathbb{F}$ with a multiplicative inverse of its zero element, denoted by 0^{-1}, is called an *augmented field.* The union of a differential field $\mathbb{F}$ with a multiplicative inverse 0^{-1} for its zero element is called an *augmented differential field.* The zero element 0 and the element 0^{-1} are then called the singular elements in that augmented (differential) field.

Remarks. (1) By convention, multiplicity and powers of the zero element 0 in a number field is insignificant. But this is not the case in an augmented (differential) field when 0^{-1} comes into play. For instance, $(0+0)0^{-1}=1$, whereas $0\cdot 0^{-1}+0\cdot 0^{-1}=2$. To avoid indefinite computation results, the following rules should be observed when the operands include the singular numbers 0 and 0^{-1}:

(i) Keep track of multiplicity and powers of the singular numbers 0 and 0^{-1} at each intermediate stage, e.g., $0+0=2\cdot 0$, $0\cdot 0=0^2$, $0^{-1}+0^{-1}=2\cdot 0^{-1}$, $0^{-1}\cdot 0^{-1}=0^{-2}$.

(ii) Combine 0 and 0^{-1} at each intermediate stage according to the rules: $0+0^{-1}=0^{-1}$, $0\cdot 0^{-1}=1$, until only one type of singular number, 0 or 0^{-1}, is left.

(iii) At the last stage, combine the singular number, 0 or 0^{-1}, that remains from the preceeding computations with a regular number C by the rules:

$$0+C=C,\quad 0\cdot C=0,\quad 0^C=0,\quad C^0=1,$$

or

$$0^{-1}+C=0^{-1},\quad 0^{-1}\cdot C=0^{-1},\quad 0^{-C}=0^{-1},\quad C^{0^{-1}}=0^{-1},$$

or interpret the final result according to specific application purposes.

(2) Let $\mathbb{F}$ be an augmented differential field of functions $f: I\to\mathbb{F}_c$ and let $f, g\in\mathbb{F}$ such that $g(t)=1/f(t)$, $t\in\boldsymbol{I}$. If $t_0\in I$ is a continuity and an isolated zero of f, then $g(t_0)=0^{-1}$. But as $t\to t_0$, $g(t_0)\to\pm\infty\neq 0^{-1}$, i.e., $g(t_0)$ is discontinuous at t_0, and consequently nondifferentiable there. Now let $J\subseteq I$ be an open subinterval of positive measure. If $f(t)\equiv 0$, $t\in J$, then $g(t)\equiv 0^{-1}$ and $g'(t)\equiv 0$, $t\in J$. In particular, for every $t_0\in J$, $g(t)\to 0^{-1}$ as $t\to t_0$, i.e., $g(t)$ is continuous on J. A fundamental solution to the equation $\dot{x}=g(t)x$, $t\in J$, is then symbolically denoted by $x=\exp(0^{-1}t)$.

(3) In what follows we shall denote by $\mathbb{R}^0(\mathbb{C}^0)$ the augmented field $\mathbb{R}\cup\{0^{-1}\}\cdot(\mathbb{C}\cup\{0^{-1}\})$.

In order to apply the notions of D-field and augmented D-field to the analysis of time-varying linear systems (1.1), it is highly desirable that if the elements $a_{ij}(t)$ of the coefficient matrix $\mathbf{A}(t)$ are in an (augmented) D-field $\mathbb{F}$, then the elements $x_{ij}(t)$ of any fundamental solution matrix $\mathbf{X}_A(t)$ for (1.1) are also in that (augmented) field

$\mathbb{F}$. This motivates the following definition which parallels the conventional notion of algebraic closedness in the theory of number fields.

DEFINITION 2.3. An (augmented) D-field $\mathbb{F}$ is said to be *differentially closed* if for any finite number of elements $a_i \in \mathbb{F}$, $i = 1, 2, \cdots, n$, the nth-order linear differential equation

$$\frac{d^n}{dt^n} y + \alpha_n(t) \frac{d^{n-1}}{dt^{n-1}} y + \cdots + \alpha_2(t) \frac{d}{dt} y + \alpha_1(t) y = 0$$

has a fundamental set of n solutions $y_i \in \mathbb{F}$.

We now define a differentially closed augmented D-field that will be used in the sequel.

DEFINITION 2.4. Let $\mathbf{C}^{\infty}_{\text{a.e.}}(I, \mathbb{F}_c)$ be the set of almost everywhere (a.e.) $\mathbf{C}^{\infty}$ functions $f: I \to \mathbb{F}_c$, $I \subseteq \mathbb{R}$, and $\mathbb{F}_c = \mathbb{R}^0$ (or $\mathbb{F}_c = \mathbb{C}^0$). Then an augmented differential field $\mathbb{F}$ with the constant subfield $\mathbb{F}_c = \mathbb{R}^0$ ($\mathbb{F}_c = \mathbb{C}^0$) is defined by the set $\mathbf{C}^{\infty}_{\text{a.e.}}(I, \mathbb{F}_c)$, together with the operations $\{+, \cdot, d/dt\}$.

Remark. The differential closedness of the augmented D-field $\mathbb{F}$ defined above can be verified by the well-known reduction of order technique for linear differential equations and by mathematical induction.

Next we construct a vector space over an (augmented) D-field.

DEFINITION 2.5. Let $\mathbb{V}(\mathbb{F})$ be an n-dimensional vector space over a field $\mathbb{F}$. Let $\boldsymbol{\beta} = \{\mathbf{b}^1, \mathbf{b}^2, \cdots, \mathbf{b}^n\}$, $\mathbf{b}^k \in \mathbb{V}$, be a basis for $\mathbb{V}$ and $\mathbf{x}_\beta = [x_1, x_2, \cdots, x_n]$, $x_k \in \mathbb{F}$, be the coordinate vector relative to $\boldsymbol{\beta}$ for each $\mathbf{x} \in \mathbb{V}$. Then $\mathbb{V}$ is called a *differential vector space*, or simply a *D-space*, if $\mathbb{F}$ adopts a derivation operator $\boldsymbol{\delta}$ so that it becomes a differential field. The derivative operation induced on $\mathbb{V}$ is given by:

(a) $\boldsymbol{\delta}\{\mathbf{x}\} = [\boldsymbol{\delta}\{x_1\}, \boldsymbol{\delta}\{x_2\}, \cdots, \boldsymbol{\delta}\{x_n\}]$,

(b) $\boldsymbol{\delta}\{\mathbf{x}+\mathbf{y}\} = \boldsymbol{\delta}\{\mathbf{x}\} + \boldsymbol{\delta}\{\mathbf{y}\}$,

(c) $\boldsymbol{\delta}\{a\mathbf{x}\} = \boldsymbol{\delta}\{a\}\mathbf{x} + a\boldsymbol{\delta}\{\mathbf{x}\}$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{V}$ and for all $a \in \mathbb{F}$. A subset $\mathbb{W}$ of $\mathbb{V}$ is called a *D-subspace* if $\mathbb{W}$ is itself a D-space over $\mathbb{F}$.

An interesting and important structure in a D-space is provided by the following definition.

DEFINITION 2.6. Let $\mathbb{V}(\mathbb{F})$ be an n-dimensional D-space over $\mathbb{F}$ with a basis $\boldsymbol{\beta} = \{\mathbf{b}^k\}$. The set of all constant linear combinations of the basis vectors $\mathbf{b}^k \in \boldsymbol{\beta}$ forms an n-dimensional vector space over the constant subfield $\mathbb{F}_c$, called a *partial space* of $\mathbb{V}$ *relative to* β and denoted by $\mathbb{V}_\beta(\mathbb{F}_c)$. In particular, if all the coordinate vectors for $\mathbf{b}^k \in \boldsymbol{\beta}$ are constants, $\mathbb{V}_\beta(\mathbb{F}_c)$ is called the *constant partial space* of $\mathbb{V}$ and is denoted by $\mathbb{V}_c$.

The adoption of the derivative operation on a D-space $\mathbb{V}$ introduces an important family of nonlinear operators on $\mathbb{V}$, defined as follows.

DEFINITION 2.7. Let $\mathscr{P}: \mathbb{V} \to \mathbb{V}$ be an operator on a D-space $\mathbb{V}$. Then $\mathscr{P}$ is said to be partially linear if

(a) $\mathscr{P}\{\mathbf{x}+\mathbf{y}\} = \mathscr{P}\{\mathbf{x}\} + \mathscr{P}\{\mathbf{y}\}$,

(b) $\mathscr{P}\{C\mathbf{x}\} = C\mathscr{P}\{\mathbf{x}\}$

for all $C \in \mathbb{F}_c$ and for all $\mathbf{x}, \mathbf{y} \in \mathbb{V}$.

It is readily verified that if $\mathscr{P}_1$ and $\mathscr{P}_2$ are partially linear on V, then $\mathscr{P}_1 + \mathscr{P}_2$ and $\mathscr{P}_1\mathscr{P}_2$ are also partially linear. In essence, a partially linear operator on $\mathbb{V}$ is a nonlinear operator that is linear on every partial space $\mathbb{V}_\beta(\mathbb{F}_c)$ in $\mathbb{V}$. In particular, every linear operator on $\mathbb{V}$ is partially linear. Note that the derivation operator $\boldsymbol{\delta}$ on $\mathbb{V}$ is partially linear. In the next section we investigate linear differential equations (1.1) viewed as partially linear operators on some D-space.

Before concluding this section, we recall the following important result on matrix decomposition, which will be needed in the subsequent sections of this paper. In the sequel, a matrix $\mathbf{A}=[a_{ij}]$, $a_{ij}\in\mathbb{F}$, $i,j=1,2,\cdots,n$, will be denoted by $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$.

DEFINTION 2.8 [15]. Let $A\in\mathbb{M}_n(\mathbb{F})$, where $\mathbb{F}$ is an (augmented) differential field with $\mathbb{F}_c=\mathbb{C}$. ($\mathbb{F}_c=\mathbb{C}^0$). The *Cartesian decomposition* of $\mathbf{A}$ is defined by $\mathbf{A}=\operatorname{Re}\{\mathbf{A}\}+i\operatorname{Im}\{\mathbf{A}\}$, where $\operatorname{Re}\{\mathbf{A}\}$ and $\operatorname{Im}\{\mathbf{A}\}$ are Hermitian, and are given by $\operatorname{Re}\{\mathbf{A}\}=(\mathbf{A}+\mathbf{A}^*)/2$, and $\operatorname{Im}\{\mathbf{A}\}=(\mathbf{A}-\mathbf{A}^*)/2i$.

3. The LDE operator on $\mathbb{V}(\mathbb{F})$. In the sequel we will be concerned with the n-dimensional vector space $\mathbb{V}(\mathbb{F})=\mathbb{F}^n$, where $\mathbb{F}$ is the differential field $\mathbb{F}=\mathbf{C}^{\infty}_{\text{a.e.}}(I,\mathbb{F}_c)$, $I=[t_0,\infty)$, and $\mathbb{F}_c=\mathbb{C}^0$. The derivative operation $d\mathbf{x}/dt$ on $\mathbb{V}$ will be denoted by $\dot{\mathbf{x}}$ (or $(\mathbf{x})'$). With this notation we define the LDE (Linear Differential Equation) operator on $\mathbb{V}$ as follows.

DEFINITION 3.1. Let $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$. The LDE *operator* $\mathscr{P}_{\mathbf{A}}:\mathbb{V}\to\mathbb{V}$ is defined by $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}(t)\}=\mathbf{A}(t)\mathbf{x}(t)-\dot{\mathbf{x}}(t)$, $t\in I$. The matrix $\mathbf{A}$ will be called the *characteristic matrix* for the LDE operator $\mathscr{P}_{\mathbf{A}}$.

Now we can establish the following properties of an LDE operator.

THEOREM 3.1. *Let* $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$. *Then*:

(a) *The* LDE *operator* $\mathscr{P}_{\mathbf{A}}$ *is partially linear*;

(b) *The mapping* $\mathscr{P}_{\mathbf{A}}:\mathbb{V}\to\mathbb{V}$ *is onto*;

(c) *The kernel* $\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ *of* $\mathscr{P}_{\mathbf{A}}$ *is an n-dimensional partial space* $\mathbb{V}_{\beta}(\mathbb{F}_c)$ *of* $\mathbb{V}$, *where* $\boldsymbol{\beta}\subseteq\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ *is a basis for* $\mathbb{V}$.

Proof. (a) This part of the proof can be verified by Definition 2.7.

(b) If suffices to show that $\mathbb{V}\subseteq\mathscr{P}_{\mathbf{A}}\{\mathbb{V}\}$. Let $\mathbf{y}\in\mathbb{V}$. We need to show that there exists an $\mathbf{x}\in\mathbb{V}$ such that $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}\}=\mathbf{A}\mathbf{x}-\dot{\mathbf{x}}=\mathbf{y}$. Recognizing $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}\}=\mathbf{y}$ as the nonhomogeneous linear differential equation $\dot{\mathbf{x}}(t)=\mathbf{A}(t)\mathbf{x}(t)-\mathbf{y}(t)$, and invoking the existence theory for linear differential equations, we have the existence of such a solution $\mathbf{x}\in\mathbb{V}$.

(c) Since $\mathbf{x}\in\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ implies that $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}\}=\mathbf{0}$, $\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ contains nothing but solutions to the homogeneous linear differential equation

$$\dot{\mathbf{x}}(t)=\mathbf{A}(t)\mathbf{x}(t). \tag{3.1}$$

From the theory of linear differential equations, $\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ is the n-dimensional solution space of equation (3.1) over $\mathbb{F}_c$. Since $\mathbf{K}(\mathscr{P}_{\mathbf{A}})\subseteq\mathbb{V}$, it is a partial space of $\mathbb{V}$ with a basis consisting of n linearly independent (over $\mathbb{F}_c$) solutions to (3.1). Now, let $\mathbf{X}_{\mathbf{A}}$ be the matrix consisting of the n basis vectors in β; then $\det\mathbf{X}_{\mathbf{A}}(t)\neq 0$, $t\in I$. Therefore, β is also a basis of $\mathbb{V}$. □

Remarks. (1) Let $\boldsymbol{\beta}=\{\mathbf{x}^1,\mathbf{x}^2,\cdots,\mathbf{x}^n\}$ be a basis for the partial space $\mathbf{K}(\mathscr{P}_{\mathbf{A}})$. Then the matrix $\mathbf{X}_{\mathbf{A}}=[\mathbf{x}^1\mathbf{x}^2\cdots\mathbf{x}^n]$ satisfies $\mathscr{P}_{\mathbf{A}}\{\mathbf{X}_{\mathbf{A}}\}=\mathbf{0}$, and is known as a fundamental solution matrix for a linear differential equation (3.1) represented by $\mathscr{P}_{\mathbf{A}}$. In the sequel, $\mathbf{X}_{\mathbf{A}}$ will be called a *fundamental matrix* associated with $\mathbf{A}$, and for $\mathscr{P}_{\mathbf{A}}$.

(2) The partial space $\mathbb{V}_{\beta}(\mathbb{F}_c)=\mathbf{K}(\mathscr{P}_{\mathbf{A}})$ and the constant partial space $\mathbb{V}_c$ are of the same dimension n and over the same field $\mathbb{F}_c$; therefore they are isomorphic. For any $\mathbf{x}_0\in\mathbb{V}_c$, $\mathscr{P}_{\mathbf{A}}\{\mathbf{X}_{\mathbf{A}}\mathbf{x}_0\}=\mathbf{0}$, i.e., $\mathbf{X}_{\mathbf{A}}\mathbf{x}_0=\mathbf{x}\in\mathbb{V}_{\beta}$. Since $\det\mathbf{X}_{\mathbf{A}}(t)\neq 0$, $t\in I$, $\mathbf{X}_{\mathbf{A}}^{-1}$ exists. Thus $\mathbf{X}_{\mathbf{A}}$ is an isomorphism that maps $\mathbb{V}_c$ onto $\mathbb{V}_{\beta}$. In particular, for any $t_0\in I$, the state-transition matrix $\boldsymbol{\phi}_{\mathbf{A}}(t,t_0)$ given by (1.2) (also known as the normalized fundamental matrix) is the isomorphism that maps any given initial vector $\mathbf{x}(t_0)=\mathbf{x}_0\in\mathbb{V}_c$ to the state evolution vector (the unique solution to (1.1)) $\mathbf{x}(t)\in\mathbb{V}_{\beta}(\mathbb{F})$.

(3) Note that not all partial spaces of $\mathbb{V}$ constitute a kernel for LDE operators. In particular, linear independence over $\mathbb{F}_c$ does not imply linear independence over $\mathbb{F}$.

The following theorem restates some well-known properties of a fundamental matrix $\mathbf{X}_{\mathbf{A}}$ for an LDE operator $\mathscr{P}_{\mathbf{A}}$.

THEOREM 3.2 [7], [22]. *Let $\mathscr{P}_{\mathbf{A}}$ be an* LDE *operator on $\mathbb{V}$ with characteristic matrix $\mathbf{A}\in\mathbb{M}_{\eta}(\mathbb{F})$. Let $\mathbf{X}_{\mathbf{A}}$ be a fundamental matrix for $\mathscr{P}_{\mathbf{A}}$. Then*:

(a) $\det \mathbf{X}_{\mathbf{A}} = \exp\int \operatorname{tr} \mathbf{A}(\tau)\, d\tau$;

(b) *If* $\mathbf{A}=\mathbf{B}+\mathbf{C}$, *then* $\mathbf{X}_{\mathbf{A}}=\mathbf{X}_{\mathbf{B}}\mathbf{X}_{\mathbf{Q}}$, *where* $\mathbf{Q}=\mathbf{X}_{\mathbf{B}}^{-1}\mathbf{C}\mathbf{X}_{\mathbf{B}}$;

(c) *If* $\mathbf{A}$ *is semiproper on* I, *then* $\mathbf{X}_{\mathbf{A}}=\exp\int \mathbf{A}(\tau)\, d\tau$.

The following result will be used in the next section.

THEOREM 3.3. *Let $\mathbf{A}, \mathbf{X}_{\mathbf{A}}\in\mathbb{M}_{\eta}(\mathbb{F})$ such that $\mathscr{P}_{\mathbf{A}}\{\mathbf{X}_{\mathbf{A}}\}=\mathbf{0}$. If $\mathbf{X}_{\mathbf{A}}$ is unitary, then $\mathbf{A}$ is skew-Hermitian. If $\mathbf{A}$ is skew-Hermitian, then there exists a unitary $\mathbf{X}_{\mathbf{A}}$.*

Proof. Suppose that $\mathbf{X}_{\mathbf{A}}$ is unitary. Then $\mathscr{P}_{\mathbf{A}}\{\mathbf{X}_{\mathbf{A}}\}=\mathbf{0}$ implies that $\mathbf{A}=(\mathbf{X}_{\mathbf{A}})'\mathbf{X}_{\mathbf{A}}^{*}$, so that $\mathbf{A}^{*}=\mathbf{X}_{\mathbf{A}}[(\mathbf{X}_{\mathbf{A}})']^{*}=\mathbf{X}_{\mathbf{A}}(\mathbf{X}_{\mathbf{A}}^{*})'$. Since $\mathbf{A}+\mathbf{A}^{*}=(\mathbf{X}_{\mathbf{A}}\mathbf{X}_{\mathbf{A}}^{*})'=\mathbf{0}$, $\mathbf{A}=-\mathbf{A}^{*}$.

Conversely, suppose that $\mathbf{A}$ is skew-Hermitian. Let $\mathbf{Y}_{\mathbf{A}}$ be an arbitrary fundamental matrix for $\mathscr{P}_{\mathbf{A}}$. Then $(\mathbf{Y}_{\mathbf{A}})'=\mathbf{A}\mathbf{Y}_{\mathbf{A}}$, and

$$(\mathbf{Y}_{\mathbf{A}}^{*})'=[(\mathbf{Y}_{\mathbf{A}})']^{*}=\mathbf{Y}_{\mathbf{A}}^{*}\mathbf{A}^{*}=-\mathbf{Y}_{\mathbf{A}}^{*}\mathbf{A}=-\mathbf{Y}_{\mathbf{A}}^{*}(\mathbf{Y}_{\mathbf{A}})'\mathbf{Y}_{\mathbf{A}}^{-1}.$$

Thus, $(\mathbf{Y}_{\mathbf{A}}^{*})'\mathbf{Y}_{\mathbf{A}}+\mathbf{Y}_{\mathbf{A}}^{*}(\mathbf{Y}_{\mathbf{A}})'=(\mathbf{Y}_{\mathbf{A}}^{*}\mathbf{Y}_{\mathbf{A}})'=0$. This implies that $\mathbf{Y}_{\mathbf{A}}^{*}\mathbf{Y}_{\mathbf{A}}=\mathbf{C}$, which is a nonsingular constant Hermitian matrix, and thus $\mathbf{C}$ can be written as $\mathbf{C}=\mathbf{B}^{*}\mathbf{B}$, for some nonsingular constant matrix $\mathbf{B}$. Now let $\mathbf{X}_{\mathbf{A}}=\mathbf{Y}_{\mathbf{A}}\mathbf{B}^{-1}$; hence $\mathbf{X}_{\mathbf{A}}^{*}\mathbf{X}_{\mathbf{A}}=(\mathbf{B}^{-1})^{*}\mathbf{Y}_{\mathbf{A}}^{*}\mathbf{Y}_{\mathbf{A}}\mathbf{B}^{-1}=\mathbf{I}$. □

Next, we introduce the notions of D-eigenvectors and D-eigenvalues for an LDE operator $\mathscr{P}_{\mathbf{A}}$, and its characteristic matrix $\mathbf{A}$, where the prefix "D" is used to avoid confusion with the ordinary eigenvectors and eigenvalues for the matrix $\mathbf{A}$. Although these "D-eigen" concepts are not explicitly used in the subsequent developments in this paper, they will be useful tools for extending the results of this paper and for addressing other related issues in linear time-varying system theory [30]–[33].

DEFINITION 3.2. A nonzero vector $\mathbf{x}\in\mathbb{V}$ is called a *D-eigenvector* of an LDE operator $\mathscr{P}_{\mathbf{A}}$, and of its characteristic matrix $\mathbf{A}$, if there exists a $\gamma\in\mathbb{F}$ such that $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}(t)\}=\mathbf{A}(t)\mathbf{x}(t)-\dot{\mathbf{x}}(t)=\gamma(t)\mathbf{x}(t)$. The scalar γ is then called the *D-eigenvalue* associated with $\mathbf{x}$.

Remarks. (1) In Definition 3.2, we have used the morpheme "eigen" primarily because of the special form $\mathscr{P}_{\mathbf{A}}\mathbf{x}=\gamma\mathbf{x}$ and operator theoretic convention. However, the entities $\mathbf{x}\in\mathbb{V}$ and $\gamma\in\mathbb{F}$ satisfying $\mathscr{P}_{\mathbf{A}}\mathbf{x}=\gamma\mathbf{x}$ for a given matrix $\mathbf{A}(t)$ are not what we would call *natural* extensions of the tradiditional "eigenconcepts" used for constant matrices, even though when $\mathbf{A}(t)\equiv\mathbf{A}$ is constant, the conventional constant eigenvalues and eigenvectors are, by Definition 3.2, D-eigenvalues and D-eigenvectors of $\mathbf{A}$. In fact, by the well-known Existence Theorem for solutions of linear systems (1.1), *any* scalar function $\gamma\in\mathbb{F}$ is a D-eigenvalue of *any* LDE operator with $\mathbf{A}\in\mathbb{M}_{n}(\mathbb{F})$ (but not every vector $\mathbf{x}\in\mathbb{V}$ is a D-eigenvector for a given LDE operator $\mathscr{P}_{\mathbf{A}}$). Therefore, the "D-eigenconcepts" as they appear in Definition 3.2 may have only limited usefulness in applications such as solving for analytical solutions of linear time-varying systems (1.1), and/or analyzing stability of linear time-varying systems (1.1). However, we have recently used the D-eigenconcepts in Definition 3.2 to develop two new notions called *ED*-eigenvectors and *ED*-eigenvalues ("*ED*-" stands for "Essential *D*-") which have been quite useful in certain theoretical and practical applications [30]–[33].

(2) In a recent paper [24] (see also [23]), Wu introduces the notions of X-eigenvectors and X-eigenvalues for arbitrary linear time-varying systems (1.1) ("X-" stands for "Extended-") and has treated those notions as extensions of the conventional eigenconcepts for linear time-invariant systems (1.1). Although the D-eigenconcepts we have introduced are derived from a totally different line of reasoning, those concepts do coincide with Wu's X-eigenconcepts when the coefficient matrix $\mathbf{A}$ in (1.1) is restricted to the D-field $\mathbb{F}$. By a similar argument used in Remark 1 above, *any* (locally

integrable) scalar function $\lambda(t)$ is an X-eigenvalue for *any* (locally integrable) matrix $\mathbf{A}(t)$. Therefore the very essence of being "eigen" is lost in such attempts to extend the traditional eigenconcepts for linear time-invariant systems (1.1). Our new results on ED-eigenvalues and ED-eigenvectors [30], [33] are believed to constitute a more precise and more meaningful basis, compared to the X-eigenconcepts, for extending traditional time-invariant eigenconcepts to time-varying linear dynamical systems (1.1).

(3) In an operator theoretic sense we can speak of the eigenvalues $\lambda(t)$ and eigenvectors $\mathbf{y}(t)$ of a matrix $\mathbf{A}(t)$. To find $\lambda(t)$ and $\mathbf{y}(t)$ we need to solve the secular equation $[\mathbf{A}(t)-\lambda(t)\mathbf{I}]\mathbf{y}(t)=0$. However, to find the D-eigenvalues $\gamma(t)$ and D-eigenvectors $\mathbf{x}(t)$ of $\mathbf{A}(t)$, we may have to solve the linear differential equation

$$[\mathbf{A}(t)-\gamma(t)\mathbf{I}]\mathbf{x}(t)=\dot{\mathbf{x}}(t),$$

which in general is not an easy matter. It should be pointed out here that the "recursive" procedure for finding X-eigenvectors and X-eigenvalues presented by Wu in [24] is, in fact, a trial-and-error procedure.

(4) In the special case $\mathbf{A}(t)=\mathbf{A}\in\mathbb{M}_n(\mathbb{F}_c)$ and $\mathbf{x}\in\mathbb{V}(\mathbb{F})$, where $\mathbb{F}=\mathbf{C}^{\infty}_{\text{a.e.}}(I,\mathbb{F}_c)$, $\mathbb{F}_c=\mathbb{R}$, the LDE operator represents a linear differential equation with a real, constant coefficient matrix $\mathbf{A}$. In this case, it is often necessary to extend the underlying field $\mathbb{F}_c$ to $\mathbb{C}$ in order to have a full set of eigenvalues and eigenvectors. However, there always exists a full set of constant D-eigenvalues $\Gamma=\{\gamma_k\}$ in $\mathbb{F}_c=\mathbb{R}$ and a full set of D-eigenvectors in $\mathbb{V}(\mathbb{F})$. Moreover, to every eigenvalue λ of $\mathbf{A}$ there corresponds a D-eigenvalue $\gamma\in\Gamma$ such that $\gamma=\operatorname{Re}\{\lambda\}$; the corresponding D-eigenvector is also real-valued, but no longer constant. Therefore, the well-known stability criterion that a linear time-invariant system is asymptotically stable if and only if all eigenvalues λ of $\mathbf{A}$ have a negative real part can be modified to read: if and only if all the D-eigenvalues $\gamma\in\Gamma$ are negative.

(5) In the special case $\mathbf{A}(t)=\mathbf{A}\in\mathbb{M}_n(\mathbb{F}_c)$ and $\mathbf{x}(t)=\mathbf{x}\in\mathbb{V}_c$, the LDE operator becomes $\mathscr{P}_{\mathbf{A}}\{\mathbf{x}\}=\mathbf{A}\mathbf{x}$, and the D-eigenvectors $\mathbf{x}$ and D-eigenvalues γ of $\mathscr{P}_{\mathbf{A}}$ and $\mathbf{A}$ are given by $\mathbf{A}\mathbf{x}=\gamma\mathbf{x}$. In other words, the (partially linear) LDE operator then coincides with the (linear) left-multiplication operator $\mathscr{L}_{\mathbf{A}}\{\mathbf{x}\}=\mathbf{A}\mathbf{x}$ on $\mathbb{V}_c$, and the notions of D-eigenvalues and D-eigenvectors then coincide with the conventional notions of eigenvalues and eigenvectors on $\mathbb{V}_c$.

4. *D*-similarity transformations. In this section we study LDE operators under a "change of coordinate" transformation on $\mathbb{V}$. Let $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be two distinct ordered bases for $\mathbb{V}$. Let $\mathbf{x}_\beta$ and $\mathbf{x}_\gamma$ be the coordinate vectors for a vector $\mathbf{x}\in\mathbb{V}$ relative to the bases $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. Then we can always find an invertible matrix $\mathbf{L}\in\mathbb{M}_n(\mathbb{F})$ such that $\mathbf{x}_\beta=\mathbf{L}\mathbf{x}_\gamma$, for all $\mathbf{x}\in\mathbb{V}$. The matrix $\mathbf{L}$ is known as a "change of coordinate" (CC) transformation matrix. Since $\mathbf{L}$ is invertible, the change of coordinate transformation is an equivalence relation. A linear operator $\mathscr{L}$ on $\mathbb{V}$ under CC transformations takes different forms, which are said to be similar.

Now let $\mathscr{P}_{\mathbf{A}}$ be an LDE operator on $\mathbb{V}$ with characteristic matrix $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$ relative to an ordered basis $\boldsymbol{\beta}$. Let $\boldsymbol{\gamma}$ be another ordered basis and $\mathbf{L}\in\mathbb{M}_n(\mathbb{F})$ be the CC transformation matrix. Then

$$\mathscr{P}_{\mathbf{A}}\{\mathbf{x}_\beta\}=\mathbf{P}_{\mathbf{A}}\{\mathbf{L}\mathbf{x}_\gamma\}=(\mathbf{A}\mathbf{L}-\dot{\mathbf{L}})\mathbf{x}_\gamma-\mathbf{L}\dot{\mathbf{x}}_\gamma=\mathbf{y}_\beta.$$

This latter result, when expressed with respect to the basis $\boldsymbol{\gamma}$, becomes

$$\mathbf{y}_\gamma=\mathbf{L}^{-1}\mathbf{y}_\beta=(\mathbf{L}^{-1}\mathbf{A}\mathbf{L}-\mathbf{L}^{-1}\dot{\mathbf{L}})\mathbf{x}_\gamma-\dot{\mathbf{x}}_\gamma=\mathbf{B}\mathbf{x}_\gamma-\dot{\mathbf{x}}_\gamma=\mathscr{P}_{\mathbf{B}}\{\mathbf{x}_\gamma\},$$

where the matrix $\mathbf{B}\in\mathbb{M}_n(\mathbb{F})$ is given by

$$\mathbf{B}=\mathbf{L}^{-1}\mathbf{A}\mathbf{L}-\mathbf{L}^{-1}\dot{\mathbf{L}}=\mathscr{T}\{\mathbf{A}\} \tag{4.1}$$

and is the characteristic matrix for the same LDE operator relative to the basis $\boldsymbol{\gamma}$. Notice that $d(\mathbf{L}\mathbf{L}^{-1})/dt = \dot{\mathbf{I}} = \mathbf{0}$. It is straightforward to verify the following result.

THEOREM 4.1. *The transformation* $\mathscr{T}: \mathbb{M}_n(\mathbb{F}) \to \mathbb{M}_n(\mathbb{F})$ *defined by* (4.1) *is an equivalence relation.*

Now we can define D-similarity transformations of LDE operators, and of their characteristic matrices, as follows.

DEFINITION 4.1. (a) The transformation $\mathscr{T}: \mathbb{M}_n(\mathbb{F}) \to \mathbb{M}_n(\mathbb{F})$ defined by (4.1) is called a *D-similarity transformation* of a matrix $\mathbf{A} \in \mathbb{M}_n(t)$ if and only if

$$\det \mathbf{L}(t) \equiv C \neq 0$$

for some constant $C \in \mathbb{F}_c$ and for all $t \in I$. The matrices $\mathbf{A}$ and $\mathbf{B}$ in (4.1) are then said to be D-similar via $\mathbf{L}$.

(b) Two LDE operators $\mathscr{P}_{\mathbf{A}}$ and $\mathscr{P}_{\mathbf{B}}$ are said to be D-similar via $\mathbf{L}$ if and only if their associated characteristic matrices $\mathbf{A}$ and $\mathbf{B}$ are D-similar via $\mathbf{L}$.

Remarks. (1) The restriction that $\det \mathbf{L}$ be a nonzero constant provides many desirable properties which will be used in subsequent developments in this paper.

(2) D-similarity transformations coincide with conventional similarity transformations on the constant partial space $\mathbb{V}_c$.

The next result relates D-similarity transformations on a matrix $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$ with the usual similarity transformations on $\mathbf{A}$ over $\mathbb{F}$. This relation may be useful in the construction of D-similarity transformations for particular applications.

THEOREM 4.2. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{L} \in \mathbb{M}_n(\mathbb{F})$ *such that* $\det \mathbf{L} \equiv C \neq 0$. *If* $\mathbf{B} = \mathbf{L}^{-1}\mathbf{A}\mathbf{L} - \mathbf{L}^{-1}\dot{\mathbf{L}}$, *then there exists a matrix* $\mathbf{C} \in \mathbb{M}_n(\mathbb{F})$ *such that* $\mathbf{B} = \mathbf{L}^{-1}(\mathbf{A} - \mathbf{C})\mathbf{L}$.

Proof. Suppose that $\mathbf{B} = \mathbf{L}^{-1}\mathbf{A}\mathbf{L} - \mathbf{L}^{-1}\dot{\mathbf{L}}$. Let $\mathbf{C} = \dot{\mathbf{L}}\mathbf{L}^{-1}$. Then $\mathbf{C} \in \mathbb{M}_n(\mathbb{F})$ and $\mathbf{B} = \mathbf{L}^{-1}(\mathbf{A} - \mathbf{C})\mathbf{L}$. □

Note that the matrix $\mathbf{C}$ defines an LDE operator $\mathscr{P}_c$ such that $\mathbf{P}_c\{\mathbf{L}\} = \mathbf{C}\mathbf{L} - \dot{\mathbf{L}} = \mathbf{0}$, or $\mathbf{L} = \mathbf{X}_c$. In essence, Theorem 4.2 states that if $\mathbf{B}$ is D-similar to $\mathbf{A}$ via $\mathbf{L}$, then $\mathbf{B}$ is also similar (in the usual sense) to a portion of $\mathbf{A}$ via $\mathbf{L}$. In particular, if $\mathbf{B}$ is *unitarily D-similar* to $\mathbf{A}$, i.e., the transformation matrix $\mathbf{L}$ is unitary (thus $|\det \mathbf{L}| = 1$), we have the following interesting corollary.

COROLLARY 4.1. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{U} \in \mathbb{M}_n(\mathbb{F})$, *where* $\mathbf{U}$ *is a unitary matrix. If* $\mathbf{B}$ *is unitarily D-similar to* $\mathbf{A}$ *via* $\mathbf{U}$, *then* $\operatorname{Re}\{\mathbf{B}\}$ *is unitarily similar (in the usual sense) to* $\operatorname{Re}\{\mathbf{A}\}$ *via the same* $\mathbf{U}$ *matrix.*

Proof. Suppose that $\mathbf{B} = \mathbf{U}^*\mathbf{A}\mathbf{U} - \mathbf{U}^*\dot{\mathbf{U}}$. By Theorem 4.2, there exists a matrix $\mathbf{C}$ such that $\dot{\mathbf{U}} = \mathbf{C}\mathbf{U}$ and $\mathbf{B} = \mathbf{U}^*(\mathbf{A} - \mathbf{C})\mathbf{U}$. Note that by Theorem 3.3, $\mathbf{C}$ is skew-Hermitian; thus,

$$\begin{aligned}
\operatorname{Re}\{\mathbf{B}\} &= (\mathbf{B} + \mathbf{B}^*)/2 \\
&= \{\mathbf{U}^*(\mathbf{A} - \mathbf{C})\mathbf{U} + [\mathbf{U}^*(\mathbf{A} - \mathbf{C})\mathbf{U}]^*\}/2 \\
&= [\mathbf{U}^*(\mathbf{A} + \mathbf{A}^*)\mathbf{U}]/2 \\
&= \mathbf{U}^* \operatorname{Re}\{\mathbf{A}\}\mathbf{U}.
\end{aligned}$$

□

In light of Theorem 3.2, we can now give another interesting characterization of a D-similarity transformation matrix.

THEOREM 4.3. *Let* $\mathbf{L} \in \mathbb{M}_n(\mathbb{F})$. *Then* $\mathbf{L}$ *is a D-similarity transformation matrix, i.e.,* $\det \mathbf{L} = C \neq 0$, *for some* $C \in \mathbb{F}_c$, *if and only if there exists a* $\mathbf{B} \in \mathbb{M}_n(\mathbb{F})$ *with* $\operatorname{tr} \mathbf{B}(t) = 0$ *such that* $\mathbf{L}(t) = \mathbf{X}_{\mathbf{B}}(t)$, *i.e.,* $\mathscr{P}_{\mathbf{B}}\{\mathbf{L}\} = \mathbf{B}\mathbf{L} - \dot{\mathbf{L}} = \mathbf{0}$.

Proof. Suppose that $\mathbf{L} \in \mathbb{M}_n(\mathbb{F})$ satisfying $\det \mathbf{L} = C \neq 0$, $C \in \mathbf{F}_c$. Let $\mathbf{B} = \dot{\mathbf{L}}\mathbf{L}^{-1}$. Then $\mathbf{B} \in \mathbb{M}_n(\mathbb{F})$ and $\mathbf{L} = \mathbf{X}_\mathbf{B}$. By Theorem 3.2,

$$\det \mathbf{L}(t) = \det \mathbf{X}_\mathbf{B}(t) = \exp \int \operatorname{tr} \mathbf{B}(t)\, dt = C \neq 0,$$

which implies that $\operatorname{tr} \mathbf{B} = 0$.

Conversely, suppose that there exists a matrix $\mathbf{B} \in \mathbb{M}_n(\mathbb{F})$ with $\operatorname{tr} \mathbf{B} = 0$, such that $\mathbf{L} = \mathbf{X}_\mathbf{B}$. Then by Theorem 3.2,

$$\det \mathbf{L}(t) = \det \mathbf{X}_\mathbf{B}(t) = \exp \int \operatorname{tr} \mathbf{B}(t)\, dt = e^s \neq 0,$$

for some constant $S \in \mathbb{F}_c$. □

At this point, it is useful to recall the following well-known result on the extent of a D-similarity transformation.

THEOREM 4.4. *Let* $\mathbf{A}, \mathbf{B}, \mathbf{L} \in \mathbb{M}_n(\mathbb{F})$. *Then* $\mathbf{B}$ *is D-similar to* $\mathbf{A}$ *via* $\mathbf{L}$, *i.e.*, $\mathbf{B} = \mathbf{L}^{-1}\mathbf{A}\mathbf{L} - \mathbf{L}^{-1}\dot{\mathbf{L}}$, *if and only if* $\mathbf{X}_\mathbf{A} = \mathbf{L}\mathbf{X}_\mathbf{B}$.

Now we define what are called *primary* D-similarity transformations. Such transformations have many important applications. For instance, they may be used to construct some desired D-similarity transformations, or, as in the main results of this paper, can be used to reduce a linear differential equation (1.1) to some simpler form. We start with the definition of the D-elementary row (column) operations on a matrix $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$.

DEFINITION 4.2. Let $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$. Any one of the following three operations on the rows (columns) of $\mathbf{A}$ is called a *D-elementary row (column) operation*:

(a) Interchanging any two rows (columns) of $\mathbf{A}$;

(b) Multiplying any row (column) of $\mathbf{A}$ by a nonzero constant $C \in \mathbb{F}_c$;

(c) Adding to any row (column) of $\mathbf{A}$ another row (column) of $\mathbf{A}$ multiplied by a scalar $a \in \mathbb{F}$.

Remark. Note that part (b) of Definition 4.2 is different from the usual definition of elementary matrix operations over a field $\mathbb{F}$. However those two definitions coincide over the constant subfield $\mathbb{F}_c$.

DEFINITION 4.3. A matrix obtained from the identity matrix $\mathbf{I}$ by application of any one of the D-elementary row (column) operations in Definition 4.2 is called a *left (right) D-elementary matrix* of type (a), (b), or (c), and will be denoted by $\mathbf{F}^a$, $\mathbf{F}^b$, or $\mathbf{F}^c$ ($\mathbf{E}^a$, $\mathbf{E}^b$, or $\mathbf{E}^c$), accordingly. The superscript ^{a}, ^{b}, or ^{c} will be replaced by ^{x} if the type of a D-elementary matrix is indeterminate.

Remarks. (1) The determinant of any D-elementary matrix is a nonzero constant. In particular, $\det \mathbf{F}^a = -1$ ($\det \mathbf{E}^a = -1$); $\det \mathbf{F}^b = C$ ($\det \mathbf{E}^b = C$), where $C \in \mathbf{F}_c$ is the nonzero constant involved in the D-elementary operation; $\det \mathbf{F}^c = 1$ ($\det \mathbf{E}^c = 1$).

(2) By Remark 1 the D-elementary matrices are always nonsingular. The inverse of $\mathbf{E}^x(\mathbf{F}^x)$ is a D-elementary matrix of the same type, and will be denoted by $\mathbf{E}^{-x}(\mathbf{F}^{-x})$.

With this notation we can now define the *primary* D-similarity transformations as follows.

DEFINITION 4.4. Let $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$. A *right (left) primary D-similarity transformation* is a D-similarity transformation with the change of coordinate matrix $\mathbf{L}$ in (4.1) given by $\mathbf{L} = \mathbf{E}^x$ ($\mathbf{L} = \mathbf{F}^{-x}$).

The first two central results of this paper (Theorems 4.5 and 4.6) can now be stated as follows.

THEOREM 4.5. *Let* $\mathbf{L}\in\mathbb{M}_n(\mathbb{F})$. *Then* $\mathbf{L}$ *is a D-similarity transformation matrix, i.e.,* $\det \mathbf{L}(t)\equiv C\neq 0$ *for some* $C\in\mathbb{F}_c$, *if and only if* $\mathbf{L}$ *can be written as a finite product of left and right D-elementary matrices.*

In order to prove this theorem, we first need to establish two important lemmas. For this purpose, denote by $\Delta_k(t)$ the kth leading principle minor of $\mathbf{L}(t)$, i.e.,

$$\Delta_k(t)=\det\begin{bmatrix} l_{11} & \cdots & l_{1k}\\ \vdots & & \vdots\\ l_{k1} & \cdots & l_{kk}\end{bmatrix}.$$

The notation $\mathbf{A}\sim\mathbf{B}$ will be used to indicate that the matrix $\mathbf{B}$ is obtained from $\mathbf{A}$ by a finite sequence of D-elementary operations.

LEMMA 4.1. *Let* $\mathbf{L}\in\mathbb{M}_n(\mathbb{F})$. *If* $\Delta_k(t)\neq 0$, $t\in I$, $k\leqq n$, *then by means of a finite number of D-elementary row operations* $\mathbf{L}(t)$ *can be reduced to a diagonal matrix* $\mathbf{D}$ *of the form*

$$\mathbf{D}(t)=\operatorname{diag}\left[\frac{\Delta_1}{\Delta_0},\frac{\Delta_2(t)}{\Delta_1(t)},\cdots,\frac{\Delta_n(t)}{\Delta_{n-1}(t)}\right],$$

where $\Delta_0=1$. *Moreover,* $\det\mathbf{D}(t)=\det\mathbf{L}(t)$.

Lemma 4.1 can be proved by induction on n and therefore the proof is omitted.

LEMMA 4.2. *Let* $\mathbf{D}\in\mathbb{M}_n(\mathbb{F})$ *be a diagonal matrix with nonzero constant determinant. Then* $\mathbf{D}$ *can be written as a product of a finite number of D-elementary matrices.*

Proof. First it will be shown, by induction on n, that $\mathbf{D}\sim\mathbf{D}_0$, where $\mathbf{D}_0=\operatorname{diag}[1,\cdots,1,\det\mathbf{D}]$. Clearly, this is true for $n=1$. Now for arbitrary n, apply the following D-elementary operations:

(a) Add the first row to the second row.

(b) Add to the first row the second row multiplied by $(1-d_{11})/d_{11}$.

(c) Add to the second row the first row multiplied by $-d_{11}$.

(d) Add to the first row the second row multiplied by $(d_{11}-1)/d_{11}^2$.

These steps lead to the final result:

$$\mathbf{D}\sim\operatorname{diag}[1,d_{11}d_{22},d_{33},\cdots,d_{\eta\eta'}]\sim\mathbf{D}_0,$$

where the last step follows from the induction hypothesis. Since $\det\mathbf{D}$ is a nonzero constant, the proof is completed with a D-elementary row operation of type (b) applied to the last row of $\mathbf{D}_0$. □

Proof of Theorem 4.5. The sufficiency is an immediate consequence of Remark 1 following Definition 4.3. To prove the necessity, suppose that $\det\mathbf{L}(t)\equiv C\neq 0$. By means of m, $m\leqq 2(n-1)$, D-elementary row and column operations we can obtain an $\mathbf{L}_1$ whose leading principal minors $\Delta_k(t)\neq 0$, $t\in I$, $k\leqq n$. Moreover, $\det\mathbf{L}_1(t)=(-1)^m C$. The proof is then completed by applying Lemma 4.1 and 4.2 to $\mathbf{L}_1$. □

Remarks. As can be seen from the proofs of Theorem 4.5 and Lemmas 4.1 and 4.2 (and from proofs of other theorems in the sequel of this paper), the D-elementary matrices of type (c) play an important role in D-similarity transformations. Thus it is appropriate to observe here some properties of a D-elementary matrix of type (c). Let $\mathbf{E}_{rs}^c$, $r\neq s$, denote a right D-elementary matrix that performs the addition of the rth column, multiplied by some nonzero scalar function $q\in\mathbb{F}$, to the sth column. Then $\mathbf{E}_{rs}^c=\mathbf{I}+\mathbf{Q}_{rs}$, where $\mathbf{I}$ is the identity matrix, and $\mathbf{Q}_{rs}=[q_{ij}]$ such that $q_{rs}=q$, and $q_{ij}=0$ otherwise. Clearly, $\dot{\mathbf{E}}_{rs}^c=\dot{\mathbf{Q}}_{rs}$. Now let $\mathbf{Z}_{rs}=[z_{ij}]$ such that $z_{ij}=0$ if $i\neq r$ or $j\neq s$. Then it can be shown that $\mathbf{Q}_{rs}\mathbf{Z}_{rs}=\mathbf{0}$ if $r\neq s$. In particular, for $r\neq s$, $\mathbf{E}_{rs}^{-c}=\mathbf{I}-\mathbf{Q}_{rs}$, and $\mathbf{E}_{rs}^{-c}\dot{\mathbf{E}}_{rs}^c=\dot{\mathbf{Q}}_{rs}$. Similarly, a left D-elementary matrix, denoted by $\mathbf{F}_{rs}^c=\mathbf{I}+\mathbf{Q}_{rs}$, adds the sth row, multiplied by the nonzero entry q_{rs} in $\mathbf{Q}_{rs}$, to the rth row. Now let $\mathbf{B}$, $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$

such that $\mathbf{B}=\mathbf{L}^{-1}\mathbf{A}\mathbf{L}-\mathbf{L}^{-1}\dot{\mathbf{L}}$, where $\mathbf{L}=\mathbf{E}_{rs}^{c}=\mathbf{I}+\mathbf{Q}_{rs}$. Then $\mathbf{B}$ can be written as $\mathbf{B}=\mathbf{E}_{rs}^{-c}\mathbf{A}\mathbf{E}_{rs}^{c}-\dot{\mathbf{Q}}_{rs}=[b_{ij}]$, where

$$
\begin{aligned}
&b_{ij}=a_{ij} && \text{if } i\neq r,\quad j\neq s,\\
&b_{rj}=a_{rj}-q_{rs}a_{sj} && \text{if } j\neq s,\\
&b_{is}=a_{is}+q_{rs}a_{ir} && \text{if } i\neq r,\\
&b_{rs}=a_{rs}+(a_{rr}-a_{ss})q_{rs}-a_{sr}q_{rs}^{2}-\dot{q}_{rs}.
\end{aligned}
\tag{4.2}
$$

The observations and notation above will be used in the proofs of Theorem 4.6 and Lemma 4.3 given below.

THEOREM 4.6. *Every matrix* $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$ *can be reduced to a diagonal matrix* $\mathbf{D}\in\mathbb{M}_n(\mathbb{F})$ *by a finite sequence of primary D-similarity transformations.*

In order to prove this theorem, the following two important lemmas are needed. The first one presented below is also a main result of this paper.

LEMMA 4.3. *Every upper-triangular matrix* $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$ *can be reduced to a diagonal matrix* $\mathbf{D}\in\mathbb{M}_n(\mathbb{F})$ *by a finite number of primary D-similarity transformations.*

Proof. Let $\mathbf{A}=[a_{ij}]\in\mathbb{M}_n(\mathbb{F})$ be an upper-triangular matrix. The proof is by induction on n. For $n=2$, let

$$
\mathbf{L}=\mathbf{E}_{12}^{c}=\mathbf{I}+\mathbf{Q}_{12}=\begin{bmatrix}1 & q_{12}\\ 0 & 1\end{bmatrix}
$$

for some $q_{12}\in F$. Now set

$$
\begin{aligned}
\mathbf{D}&=\mathbf{L}^{-1}\mathbf{A}\mathbf{L}-\mathbf{L}^{-1}\dot{\mathbf{L}}\\
&=\begin{bmatrix}a_{11} & a_{12}+(a_{11}-a_{22})q_{12}-\dot{q}_{12}\\ 0 & a_{22}\end{bmatrix}=\begin{bmatrix}a_{11} & 0\\ 0 & a_{22}\end{bmatrix}.
\end{aligned}
$$

It can be seen that q_{12} satisfies

$$
\dot{q}_{12}(t)=[a_{11}(t)-a_{22}(t)]q_{12}(t)+a_{12}(t).
$$

From classical linear differential equation theory, such a $q_{12}\in\mathbb{F}$ can be expressed as

$$
q_{12}(t)=\phi_{12}(t)\int\phi_{12}^{-1}(t)a_{12}(t)\,dt,
$$

where

$$
\phi_{12}(t)=\exp\int[a_{11}(t)-a_{22}(t)]\,dt.
$$

Now for arbitrary n, assume that Lemma 4.3 is true for $n-1$. Partition $\mathbf{A}$ into block matrices:

$$
\mathbf{A}=\left[\begin{array}{c|c}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \hline \mathbf{A}_{21} & \mathbf{A}_{22}\end{array}\right]
$$

such that $\mathbf{A}_{11}=[a_{11}]$. Let $\mathbf{D}_{22}=\operatorname{diag}[a_{22},a_{33},\cdots,a_{nn}]$, and let $\mathbf{L}_1=\operatorname{diag}[\mathbf{L}_{11},\mathbf{L}_{22}]$ such

that $\mathbf{L}_{11}=[1]$ and $\mathbf{L}_{22}^{-1}\mathbf{A}_{22}\mathbf{L}_{22}-\mathbf{L}_{22}^{-1}\dot{\mathbf{L}}_{22}=\mathbf{D}_{22}$. Then we have

$$\mathbf{B}=\mathbf{L}_1^{-1}\mathbf{A}\mathbf{L}_1-\mathbf{L}_1^{-1}\dot{\mathbf{L}}_1$$

$$=\left[\begin{array}{c|c} a_{11} & \mathbf{A}_{12}\mathbf{L}_{22} \\ \hline \mathbf{0} & \mathbf{D}_{22} \end{array}\right]=\left[\begin{array}{c|ccc} a_{11} & b_{12} & \cdots & b_{1n} \\ \hline 0 & a_{22} & 0 \cdots & 0 \\ \vdots & 0 & \ddots \quad \ddots & \vdots \\ \vdots & \vdots & \ddots \quad \ddots & 0 \\ 0 & 0 & \cdots \quad 0 & a_{nn} \end{array}\right].$$

Now let $\mathbf{L}_2=\mathbf{E}_{12}^c\mathbf{E}_{13}^c\cdots\mathbf{E}_{1n}^c$, where $\mathbf{E}_{1k}^c=\mathbf{I}+\mathbf{Q}_{1k}$. Set

$$\mathbf{D}=\mathbf{L}_2^{-1}\mathbf{A}\mathbf{L}_2-\mathbf{L}_2^{-1}\dot{\mathbf{L}}_2$$

$$=\left[\begin{array}{c|ccc} a_{11} & d_{12} & \cdots & d_{1n} \\ \hline 0 & a_{22} & 0 \cdots & 0 \\ \vdots & 0 & \ddots \quad \ddots & \vdots \\ \vdots & \vdots & \ddots \quad \ddots & 0 \\ 0 & 0 & \cdots \quad 0 & a_{nn} \end{array}\right]=\left[\begin{array}{c|ccc} a_{11} & 0 & \cdots & 0 \\ \hline 0 & a_{22} & 0 \cdots & 0 \\ \vdots & 0 & \ddots \quad \ddots & \vdots \\ \vdots & \vdots & \ddots \quad \ddots & 0 \\ 0 & 0 & \cdots \quad 0 & a_{nn} \end{array}\right],$$

where

$$d_{1k}=(a_{11}-a_{kk})q_{1k}+b_{1k}-\dot{q}_{1k}=0, \tag{4.3}$$

$k=2,3,\cdots,n$. Thus a $q_{1k}\in\mathbb{F}$ can be found as

$$q_{1k}(t)=\phi_{1k}(t)\int\phi_{1k}^{-1}(t)b_{1k}(t)\,dt, \tag{4.4}$$

where

$$\phi_{1k}(t)=\exp\int[a_{11}(t)-a_{kk}(t)]\,dt. \tag{4.5}$$

Now let $\mathbf{L}=\mathbf{L}_1\mathbf{L}_2$; the lemma then follows from Theorem 4.1. □

Lemma 4.3 has an immediate corollary, which greatly increases its applicability.

COROLLARY 4.2. *Every virtually triangular matrix* $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$ *can be reduced to a diagonal matrix* $\mathbf{D}\in\mathbb{M}_n(\mathbb{F})$ *by a finite number of primary D-similarity transformations.*

The next lemma is due to Perron [18] (see also [13] and [19]).

LEMMA 4.4. *Let* $\mathbf{A}\in\mathbb{M}_n(\mathbb{F})$. *If* $\mathbf{X}_\mathbf{A}$ *is a fundamental matrix for* $\mathscr{P}_\mathbf{A}$, *then there exist a unitary matrix* $\mathbf{U}\in\mathbb{M}_n(\mathbb{F})$ *and a nonsingular upper-triangular matrix* $\mathbf{T}\in\mathbb{M}_n(\mathbb{F})$ *such that* $\mathbf{X}_\mathbf{A}=\mathbf{U}\mathbf{T}$.

Proof of Theorem 4.6. Let $\mathbf{A}, \mathbf{X}_\mathbf{A}, \mathbf{U}$, and $\mathbf{T}$ be matrices as given in Lemma 4.4. Let $\mathbf{B}=\mathbf{T}^{-1}\dot{\mathbf{T}}$. Then $\mathbf{B}$ is upper triangular and $\mathbf{X}_\mathbf{B}=\mathbf{T}$. By Theorem 4.4, $\mathbf{B}=\mathbf{U}^{-1}\mathbf{A}\mathbf{U}-\mathbf{U}^{-1}\dot{\mathbf{U}}$. Now by Lemma 4.3, there exists a D-similarity transformation matrix $\mathbf{L}_1\in\mathbb{M}_n(\mathbb{F})$ such that $\mathbf{L}_1^{-1}\mathbf{B}\mathbf{L}_1-\mathbf{L}_1^{-1}\dot{\mathbf{L}}_1=\mathbf{D}_1$, where $\mathbf{D}_1$ is a diagonal matrix. Since $|\det\mathbf{U}|=1$, we can, taking the principal branch, define $\xi(t)=\det\mathbf{U}(t)^{-1/n}$. Now let $\mathbf{L}=\xi\mathbf{U}\mathbf{L}_1$. Then $\det\mathbf{L}=\det\mathbf{L}_1$, and $\mathbf{L}^{-1}\mathbf{A}\mathbf{L}-\mathbf{L}^{-1}\dot{\mathbf{L}}=\mathbf{D}_1-\xi^{-1}\dot{\xi}\mathbf{I}=\mathbf{D}$ is a diagonal matrix. The theorem hence follows from Theorems 4.5 and 4.1. □

Remarks. Theorem 4.6 is an exciting theoretical result and has immediate applications in the semiproper reducible systems (1.1) defined by Definition 1.2. However, there are limitations to the practical application of this technique because procedures for finding the primary D-similarity transformations for the reduction of linear systems

(1.1) with coefficient matrices $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$ involve, in general, solving systems of Riccati-type nonlinear differential equations that are usually as difficult to solve as the original ones. These limitations are significantly reduced by the following important result.

THEOREM 4.7. *For every matrix* $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$, *there exist a unitary matrix* $\mathbf{U} \in \mathbb{M}_n(\mathbb{F})$, *and a Hermitian matrix* $\mathbf{H} \in \mathbb{M}_n(\mathbb{F})$, *such that* $\mathbf{H} = \mathbf{U}^*\mathbf{A}\mathbf{U} - \mathbf{U}^*\dot{\mathbf{U}} = \mathbf{U}^* \operatorname{Re}\{\mathbf{A}\}\mathbf{U}$.

Proof. Let $\mathbf{A} \in \mathbb{M}_n(\mathbb{F})$. For convenience, let $\mathbf{A}_R = \operatorname{Re}\{\mathbf{A}\}$, and $\mathbf{A}_I = i \operatorname{Im}\{\mathbf{A}\}$. Since $\mathbf{A}_I$ is skew-Hermitian, by Theorem 3.3, there exists a unitary matrix $\mathbf{U}$ such that $\dot{\mathbf{U}} = \mathbf{A}_I\mathbf{U}$. Thus, $\mathbf{H} = \mathbf{U}^*\mathbf{A}\mathbf{U} - \mathbf{U}^*\dot{\mathbf{U}} = \mathbf{U}^*(\mathbf{A} - \mathbf{A}_I)\mathbf{U} = \mathbf{U}^*\mathbf{A}_R\mathbf{U}$. Note that $\mathbf{A}_R$ is Hermitian, and it is readily verified that $\mathbf{H}$ is also Hermitian. □

Remark. The importance of Theorem 4.7 lies in the fact that instead of investigating D-similarity transformations for a *general* matrix $\mathbf{A}$, we need to study such transformations only for the two *normal* matrices $\mathbf{A}_I$ and $\mathbf{H} = \mathbf{U}^*\mathbf{A}_R\mathbf{U}$, where $\mathbf{U}$ satisfies $\dot{\mathbf{U}} = \mathbf{A}_I\mathbf{U}$. That is, a procedure for finding D-similarity transformations for normal matrices is, in general, all that is needed. This result, when applied to the reduction of linear systems (1.1), enables us to "decompose" a general system (1.1) with coefficient matrix $\mathbf{A}(t)$ into two normal systems, i.e., a skew-Hermitian system defined by $\mathbf{A}_I(t)$, and a Hermitian system defined by $\mathbf{H}(t)$. The solution of the original system can then be expressed, by Theorem 3.2, as $\mathbf{X}_\mathbf{A} = \mathbf{U}\mathbf{X}_\mathbf{H}$. Therefore it suffices to investigate only the reduction problem for normal systems.

5. Reducible systems and reducibility. Some examples. In this section, the results obtained in the preceding section are applied to semiproper reducible linear dynamical systems (1.1) defined by Definition 1.2. Note that although the matrices $\mathbf{A}(t)$ considered in the previous section are assumed to be of class $\mathbf{C}^\infty$ almost everywhere on an interval I, that requirement is not essential when we are interested in solving a particular case of (1.1). Namely, from the proof of Theorem 4.6 it can be seen that, in a specific problem, $\mathbf{A}(t)$ is only required to be locally Lebesgue integrable. Moreover, if $\mathbf{A}(t)$ is upper-triangular, it is only necessary that the D-similarity transformation matrix $\mathbf{L}(t)$ be over a ring of finitely differentiable functions. In this sense, therefore, we can relax our results obtained over a differential field to what we called "well-defined" matrices in Definition 1.2 and Theorems 1.1 and 1.3. With this remark, Theorems 1.1–1.3 follow immediately from Theorems 4.5 and 4.6. We are now in a position to establish the following fundamental result of this paper which characterizes the solutions of semiproper reducible systems (1.1).

THEOREM 5.1. *A system* (1.1) *is semiproper reducible if and only if for any* $t_0 \in I$ *and* $\mathbf{x}(t_0) \in \mathbb{V}_c$, *the solution to* (1.1) *can be written as*

$$\mathbf{x}(t) = \mathbf{L}(t) \exp\left[\int_{t_0}^{t} \mathbf{B}(\tau)\, d\tau\right] \mathbf{L}^{-1}(t_0)\mathbf{x}(t_0) \tag{5.1}$$

for some sufficiently differentiable matrix function $\mathbf{L}(t)$ *with nonzero constant determinant and for some locally integrable matrix* $\mathbf{B}(t)$.

Proof. Suppose that there exist a locally integrable $\mathbf{B}(t)$ and a sufficiently differentiable $\mathbf{L}(t)$ with nonzero constant determinant such that (5.1) is satisfied. Let $\mathbf{y}(t) = \mathbf{L}^{-1}(t)\mathbf{x}(t)$; then

$$\mathbf{x}(t) = \mathbf{L}(t)\mathbf{y}(t). \tag{5.2}$$

Substituting (5.2) into (1.1) and (5.1) yields

$$\begin{aligned} \dot{\mathbf{y}}(t) &= [\mathbf{L}^{-1}(t)\mathbf{A}(t)\mathbf{L}(t) - \mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t)]\mathbf{y}(t) \\ &= \mathbf{B}_0(t)\mathbf{y}(t), \end{aligned} \tag{5.3}$$

$$\mathbf{y}(t_0) = \mathbf{L}^{-1}(t_0)\mathbf{x}(t_0), \qquad t_0 \in I, \tag{5.4}$$

$$
\begin{aligned}
y(t) &= \left[\exp \int_{t_0}^{t} \mathbf{B}(\tau)\, d\tau\right] \mathbf{y}(t_0) \\
&= \left[\sum_{k=0}^{\infty} \frac{1}{k!}\left[\int_{t_0}^{t} \mathbf{B}(\tau)\, d\tau\right]^k\right] \mathbf{y}(t_0).
\end{aligned} \tag{5.5}
$$

Note that the power series in (5.5) is uniformly convergent. Thus, differentiating (5.5) and comparing with (5.3) term by term yields

$$
\begin{aligned}
&\mathbf{B}_0(t) = \mathbf{B}(t), \\
&\mathbf{B}(t) \int_{t_0}^{t} \mathbf{B}(\tau)\, d\tau = \int_{t_0}^{t} \mathbf{B}(\tau)\, d\tau\, \mathbf{B}(t),
\end{aligned}
$$

for any $t_0 \in I$. Therefore, by a result of Martin [16, Thm. 1] we conclude that $\mathbf{B}(t)$ is semiproper, and consequently the given system (1.1) is semiproper reducible.

The converse is a straightforward consequence of Theorems 4.4 and 3.2. □

The remark following Definition 1.2 regarding the relationship between Lyapunov reducibility and the semiproper reducibility can now be formally stated and proven as follows.

THEOREM 5.2. *Every Lyapunov reducible system* (1.1) *is semiproper reducible.*

Proof. Suppose that system (1.1) is Lyapunov reducible and $\mathbf{L}(t)$ and $\mathbf{B}$ are as given in Definition 1.1. Since $0 < N < |\det \mathbf{L}(t)|$, we can by taking the principal branch, define $\xi(t) = (\det \mathbf{L}(t))^{1/n}$. Now let $\mathbf{L}_0(t) = \xi^{-1}(t)\mathbf{L}(t)$, then $\det \mathbf{L}_0(t) = 1$. Thus $\mathbf{L}_0(t)$ is a D-similarity transformation matrix. Moreover, let $\beta(t) = \dot{\xi}(t)/\xi(t)$; then

$$
\begin{aligned}
\mathbf{B}(t) &= \mathbf{L}_0^{-1}(t)\mathbf{A}(t)\mathbf{L}_0(t) - \mathbf{L}_0^{-1}(t)\dot{\mathbf{L}}_0(t) \\
&= \mathbf{L}^{-1}(t)\mathbf{A}(t)\mathbf{L}(t) - \mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t) + \beta(t)\mathbf{I} \\
&= \mathbf{B} + \beta(t)\mathbf{I},
\end{aligned}
$$

which is readily verified to be semiproper. □

Some illustrative examples. In the remainder of this section we work some specific examples to illustrate the application and limitations of our results.

Example 5.1. Consider the system (1.1) with an upper-triangular $\mathbf{A}(t)$ given by

$$
\mathbf{A}(t) = \begin{bmatrix}
2t & -(2t+\cos t)\, e^{-\sin t - \cos t} & (\cos t + \sin t)\, e^{-\cos t - e^{-t}} \\
0 & -\sin t & (\cos t + \sin t)\, e^{\sin t - e^{-t}} \\
0 & 0 & -e^{-t}
\end{bmatrix}.
$$

According to the inductive proof of Lemma 4.2, let

$$
\mathbf{L}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & l_{23} \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{L}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -l_{23} \\ 0 & 0 & 1 \end{bmatrix}, \quad \dot{\mathbf{L}}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \dot{l}_{23} \\ 0 & 0 & 0 \end{bmatrix},
$$

where $l_{23}(t)$ can be bound by formulas (4.4) and (4.5) as

$$
l_{23}(t) = \phi_{23}(t) \int \phi_{23}^{-1}(t) a_{23}(t)\, dt = e^{\sin t - e^{-t}},
$$

where

$$
\phi_{23}(t) = \exp\left\{\int [a_{22}(t) - a_{33}(t)]\, dt\right\} = e^{\cos t - e^{-t}};
$$

then

$$\mathbf{B} = \mathbf{L}_1^{-1}\mathbf{A}\mathbf{L}_1 - \mathbf{L}_1^{-1}\dot{\mathbf{L}}_1 = \left[\begin{array}{c|cc} a_{11} & a_{12} & a_{13}+l_{23}a_{12} \\ \hline 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{array}\right].$$

Now let

$$\mathbf{L}_2 = \begin{bmatrix} 1 & l_{12} & l_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{L}_2^{-1} = \begin{bmatrix} 1 & -l_{12} & -l_{13} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \dot{\mathbf{L}}_2 = \begin{bmatrix} 0 & \dot{l}_{12} & \dot{l}_{13} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where l_{12} and l_{13} are found to be

$$l_{12}(t) = \phi_{12}(t) \int \phi_{12}^{-1}(t) b_{12}(t)\, dt = e^{-\sin t - \cos t},$$

$$l_{13}(t) = \phi_{13}(t) \int \phi_{13}^{-1}(t) b_{13}(t)\, dt = e^{-\cos t - e^{-t}},$$

where

$$\phi_{12}(t) = \exp\left\{\int [a_{11}(t) - a_{22}(t)]\, dt\right\} = e^{t^2 - \cos t},$$

$$\phi_{13}(t) = \exp\left\{\int [a_{11}(t) - a_{33}(t)]\, dt\right\} = e^{t^2 - e^{-t}}.$$

Then

$$\begin{aligned} \mathbf{D}(t) &= \mathbf{L}_2^{-1}(t)\mathbf{B}(t)\mathbf{L}_2(t) - \mathbf{L}_2^{-1}(t)\dot{\mathbf{L}}_2(t) \\ &= \operatorname{diag}[a_{11}(t), a_{22}(t), a_{33}(t)] \\ &= \operatorname{diag}[2t, -\sin(t), -e^{-t}]. \end{aligned}$$

Now let $\mathbf{L} = \mathbf{L}_1\mathbf{L}_2$; then

$$\mathbf{L} = \begin{bmatrix} 1 & l_{12} & l_{13} \\ 0 & 1 & l_{23} \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{L}^{-1} = \begin{bmatrix} 1 & -l_{12} & l_{12}l_{23} - l_{13} \\ 0 & 1 & -l_{13} \\ 0 & 0 & 1 \end{bmatrix}, \quad \dot{\mathbf{L}} = \begin{bmatrix} 0 & \dot{l}_{12} & \dot{l}_{13} \\ 0 & 0 & \dot{l}_{23} \\ 0 & 0 & 0 \end{bmatrix}.$$

We see that $\mathbf{L}^{-1}\mathbf{A}\mathbf{L} - \mathbf{L}^{-1}\dot{\mathbf{L}} = \mathbf{D}$. Denote by $\mathbf{X}_\mathbf{D}$ a fundamental matrix associated with D; then

$$\begin{aligned} \mathbf{X}_\mathbf{D}(t) &= \operatorname{diag}\left[\exp\int d_{11}(t)\, dt, \exp\int d_{22}(t)\, dt, \exp\int d_{33}(t)\, dt\right] \\ &= \operatorname{diag}[e^{t^2}, e^{\cos t}, e^{e^{-t}}]. \end{aligned}$$

By Theorem 4.4, a fundamental matrix $\mathbf{X}_\mathbf{A}$ for (1.1) with the given $\mathbf{A}(t)$ is found to be

$$\begin{aligned} \mathbf{X}_\mathbf{A}(t) &= \mathbf{L}(t)\mathbf{X}_\mathbf{D}(t) \\ &= \begin{bmatrix} e^{t^2} & e^{-\sin t} & e^{-\cos t} \\ 0 & e^{\cos t} & e^{\sin t} \\ 0 & 0 & e^{e^{-t}} \end{bmatrix}. \end{aligned}$$

As a special case of Lemma 4.2, every linear system (1.1) with a Jordan coefficient matrix $\mathbf{A}(t)=\mathbf{J}(t)$ can be reduced to a diagonal system by primary D-similarity transformations. The next example demonstrates this idea with a Jordan block of order 3.

Example 5.2. Consider the Jordan block of order 3:

$$\mathbf{J}(t)=\begin{bmatrix}\lambda(t) & 1 & 0\\ 0 & \lambda(t) & 1\\ 0 & 0 & \lambda(t)\end{bmatrix}.$$

By the technique illustrated in Example 5.1, we can find a D-similarity transformation $\mathbf{L}$ such that

$$\mathbf{L}(t)=\begin{bmatrix}1 & t & t^2/2\\ 0 & 1 & t\\ 0 & 0 & 1\end{bmatrix},\quad \mathbf{L}^{-1}(t)=\begin{bmatrix}1 & -t & t^2/2\\ 0 & 1 & -t\\ 0 & 0 & 1\end{bmatrix},\quad \dot{\mathbf{L}}_2(t)=\begin{bmatrix}0 & 1 & t\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}.$$

Then

$$\begin{aligned}\mathbf{D}(t)&=\mathbf{L}^{-1}(t)\mathbf{B}(t)\mathbf{L}(t)-\mathbf{L}^{-1}(t)\dot{\mathbf{L}}(t)\\ &=\operatorname{diag}[\lambda(t),\lambda(t),\lambda(t)].\end{aligned}$$

The following example demonstrates a limitation of the primary D-similarity transformation technique for the reduction of linear systems (1.1).

Example 5.3. Consider a general second-order linear system (1.1) with $\mathbf{A}=[a_{ij}]\in \mathbb{M}_n(\mathbb{F})$. Suppose we attempt to reduce $\mathbf{A}(t)$ to an upper-triangular matrix by letting

$$\mathbf{L}_1=\begin{bmatrix}1 & 0\\ l_{21} & 1\end{bmatrix}.$$

By formula (4.2), in the matrix $\mathbf{B}=\mathbf{L}_1^{-1}\mathbf{A}\mathbf{L}_1-\mathbf{L}_1^{-1}\dot{\mathbf{L}}_1=[b_{ij}]$ we should set

$$b_{21}=a_{21}+(a_{22}-a_{11})l_{21}-a_{12}l_{21}^2-\dot{l}_{21}=0.$$

However, this latter condition yields the following nonlinear Riccati equation with time-varying coefficients $a_{ij}(t)$:

$$\dot{l}_{21}=a_{21}+(a_{22}-a_{11})l_{21}-a_{12}l_{21}^2,$$

which may be as difficult to solve as the original equation.

6. Summary and conclusions. A century ago Lyapunov introduced the notion of reducing a linear differential equation (1.1) with a "time-varying" coefficient matrix $\mathbf{A}(t)$ to one with a constant coefficient matrix. Since that time a great deal of research effort has been devoted to the qualitative analysis of reducible systems and procedures for construction of the reduction transformations. As a consequence, it appears that today there is not much room for further developments within the original confines of Lyapunov's reduction technique.

Inspired by recent results [26]–[29], [34], [35] for obtaining explicit solutions for time-varying semiproper systems (1.1), the present paper has formally extended the family of Lyapunov reducible systems to include those systems that can be reduced to semiproper ones via D-similarity transformations. We call such systems "semiproper reducible." Within this new framework for reducibility studies, we have defined primary D-similarity transformations, and have proven that every "well-defined" linear system (1.1) is semiproper reducible by a finite sequence of primary D-similarity transformations. We have also presented an explicit technique for constructing such transformations for systems (1.1) with virtually triangular coefficient matrices.

From the results in [26]–[29], and in the present paper, the family of time-varying linear systems which is now analytically solvable consists of all systems (1.1) with virtually triangular coefficient matrices $\mathbf{A}(t) = \mathbf{L}\mathbf{T}(t)\mathbf{L}^{-1}$. This family includes the special cases of (i) *time-invariant*, (ii) *time-varying proper*, and (iii) *time-varying semiproper* systems (1.1) and can be summarized as follows:

$$\left\{\begin{array}{c}\text{Constant}\\ \text{coefficient}\\ \mathbf{A}(t)=\mathbf{A}\end{array}\right\} \subset \left\{\begin{array}{c}\text{Proper}\\ \mathbf{A}(t)=f(t,\mathbf{A})\end{array}\right\} \subset \left\{\begin{array}{c}\text{Semiproper}\\ \mathbf{A}(t)\mathbf{A}(\tau)=\mathbf{A}(\tau)\mathbf{A}(t)\end{array}\right\} \subset \left\{\begin{array}{c}\text{Virtually triangular}\\ \text{coefficient}\\ \mathbf{A}(t)=\mathbf{L}\mathbf{T}(t)\mathbf{L}^{-1}\end{array}\right\}.$$

The last inclusion follows from a result due to Martin [16], which states that every semiproper (functionally commutative) matrix is virtually triangular. Now, by virtue of Theorem 4.7, the reduction of a general linear system (1.1) can be achieved by the reduction of two normal subsystems. Therefore, *the only family of linear systems* (1.1) *that needs further investigations, in terms of solvability, is that of the normal systems* (1.1).

As a byproduct of achieving our main results (Theorems 1.1–1.3, 4.5–4.7, 5.1, 5.2, and Lemma 4.3 and its corollary), we have also obtained some important and interesting new results in the theory of matrices over a differential field, among which are the notions of partial spaces and partially linear operators; the introduction of LDE operators and associated D-eigenvalues, D-eigenvectors, and D-similarity transformations; and Theorems 3.1, 3.3, 4.2, and 4.3, Lemmas 4.1 and 4.2, and Corollary 4.1.

Example 5.3 in § 5 of this paper has been presented to illustrate a limitation of the proposed technique for constructing D-similarity transformations for reduction of general linear systems (1.1). However, other approaches to such problems are currently being investigated, and these alternative approaches may circumvent those limitations. Further characterizations of D-similarity transformations and qualitative properties of the solutions of semiproper reducible systems are other topics currently being investigated. Results from these investigations will appear in forthcoming papers.

Acknowledgment. The authors thank the reviewer for calling their attention to the important references [23], [24] by M.-Y. Wu.

REFERENCES

[1] R. BELLMAN, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1960.
[2] H. D'ANGELO, *Linear Time-Varying Systems: Analysis and Synthesis*, Allyn and Bacon, Boston, 1970.
[3] N. P. ERUGIN, *Privodimyye sistemy (Reduced system)*, Trudy Fiz.-Mat. Inst. V. A. Steklova, XIII (1946).
[4] N. P. ERUGIN, *Linear Systems of Ordinary Differential Equations with Periodic and Quasi-Periodic Coefficients*, Academic Press, New York, 1966.
[5] G. FLOQUET, *Sur les equations differentielles linéaires à coefficients périodiques*, Ann. Sci. Ecole Norm. Sup. (4), 12 (1983), pp. 47–82.
[6] S. H. FRIEDBERG, A. J. INSEL, AND L. E. SPENCE, *Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
[7] F. R. GANTMACHER, *The Theory of Matrices*, Chelsea, New York, 1959.
[8] J. K. HALE, *Ordinary Differential Equations*, John Wiley, New York, 1969.
[9] P. HARTMAN, *Ordinary Differential Equations*, John Wiley, New York, 1964.
[10] K. KNOPP, *Theory of Functions* (F. Bagemihl, trans.), Dover, New York, 1945.
[11] E. R. KOLCHIN, *Differential Algebra and Algebraic Groups*, Academic Press, New York, 1973.
[12] P. LANCASTER AND M. TISMENETSKY, *The Theory of Matrices*, Academic Press, New York, 1985.
[13] S. LEFSCHETZ, *Differential Equations: Geometric Theory*, John Wiley, New York, 1963.
[14] A. M. LYAPUNOV, *Stability of Motion*, Academic Press, New York, 1966.
[15] M. MARCUS AND H. A. MINC, *Survey of Matrix Theory and Matrix Inequalities*, Prindle, Weber, and Schmidt, Boston, 1964.
[16] J. F. P. MARTIN, *Some results on matrices that commute with their derivatives*, SIAM. J. Appl. Math., 15 (1967), pp. 1171–1183.

[17] V. V. NEMYTSKII AND V. V. STEPANOV, *Qualitative Theory of Differential Equations*, Princeton University Press, Princeton, NJ, 1960.
[18] O. PERRON, *Über eine Matrixtransformation*, Math. Zeit., 32 (1930), pp. 456–473.
[19] W. T. REID, *Remarks on a matrix transformation for linear differential equations*, Proc. Amer. Math. Soc., 8 (1957), pp. 708–712.
[20] N. J. ROSE, *On the eigenvalues of a matrix which commutes with its derivative*, Proc. Amer. Math. Soc., 16 (1965), pp. 752–754.
[21] R. L. WHEEDEN AND A. ZYGMUND, *Measure and Integral*, Marcel Dekker, New York, 1977.
[22] D. M. WIBURG, *State Space and Linear Systems*, McGraw-Hill, New York, 1971.
[23] M.-Y. WU, *A new concept of eigenvalues and eigenvectors and its applications*, IEEE Trans. Automat. Control, AC-25 (1987), pp. 824–826.
[24] ———, *On stability of linear time-varying systems*, Internat. J. Systems Sci., 15 (1984), pp. 137–150.
[25] J. ZHU, *On linear differential equations with functionally commutative coefficient matrices*, Master's thesis, University of Alabama, Huntsville, AL, 1986.
[26] J. ZHU AND C. D. JOHNSON, *A closed-form analytic solution for the state-transition matrix of linear time-varying semiproper systems*, in Proc. SSST87, Clemson University, Clemson, SC, March 1987.
[27] ———, *Stability criteria for linear time-varying semiproper systems*, in Proc. SSST87, Clemson University, Clemson, SC, March 1987.
[28] J. ZHU AND C. H. MORALES, *Spatial decomposition of functionally commutative matrices*, submitted.
[29] ———, *On linear ordinary differential equations with functionally commutative coefficient matrices*, submitted, 1987.
[30] J. ZHU AND C. D. JOHNSON, *Unified canonical forms for linear time-varying dynamical systems*, submitted, 1988.
[31] ———, *Invariants of linear time-varying dynamical systems under D-similarity transformations*, submitted.
[32] ———, *New results for the stability analysis of time-varying linear systems*; *Part* II: *The case of reducible systems*, to appear.
[33] ———, *New results on eigenvalue concepts for linear time-varying dynamical systems*, in Proc. 26th Allerton Conference on Communication, Control, and Computing, Champaign, IL, September 1988.
[34] ———, *New results for the stability analysis of time-varying linear systems*; *part* I: *the case of reduced systems*, in Proc. ACC88, Atlanta, GA, June 1988.
[35] ———, *A necessary and sufficient criterion for semiproper linear time-varying systems*, in Proc. SSST88, Charlotte, NC, March 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 495-510, May 1989

003

GENERALIZED REACHABILITY SUBSPACES FOR SINGULAR SYSTEMS*

K. ÖZCALDIRAN† AND F. L. LEWIS‡

Abstract. One of the most important concepts of the geometric theory for proper linear systems is that of the controllability subspaces of (A, B). In an attempt to extend this concept to singular systems, it became clear that the nonequivalence of reachability and controllability for singular systems makes it necessary to define both controllability and reachability subspaces of (E, A, B) as different, though naturally related, entities.

A subspace $\mathbf{R}$ (respectively, $\mathbf{C}$) is defined to be a *generalized reachability subspace* (respectively, *generalized controllability subspace*) if there exist linear maps F and G, such that $\mathbf{R}$ (respectively, $\mathbf{C}$) is the reachable (respectively, controllable) subspace of a regular singular system $E\dot{x} = (A + BF)x + BGv$. It is proved that (1) $\mathbf{R}$ is a generalized reachability subspace if and only if it is a (generalized) $(A, E, \mathbf{B})$-invariant almost reachability subspace; (2) every subspace $\mathbf{K}$ contains a unique supremal generalized reachability subspace $\mathbf{R}^*(\mathbf{K})$; and (3) a subspace $\mathbf{C} \subset \mathbf{Im}\, \mathbf{E}$ is a generalized controllability subspace if and only if $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{K})$. In the case where $\mathbf{R}$ is a generalized reachability subspace, the F and G matrices that make $\mathbf{R}$ the reachable subspace of the closed-loop system are constructed explicitly. Spectral assignability properties of generalized reachability and controllability subspaces are also treated. For completeness, all results are finally extended to the nonregular case (i.e., $\det(sE - A) = 0$).

Key words. singular systems, descriptor systems, geometrical system theory, reachability subspaces, controllability subspaces

AMS(MOS) subject classifications. 93C05, 93C35

1. Introduction.

In recent years, the study of generalized or singular systems of the form $E\dot{x} = Ax + Bu$, with E generally singular, has become of interest (see references). Indeed, the study of singular pencils of matrices of the form $sE - A$ has been of mathematical interest since the writings of Weierstrass in 1867 [25] and Kronecker in 1890 [6]. (See [5] and [26] for a summary of their work.)

There are many reasons for the current revival of singular systems. They arise more naturally than state-space systems in the study of naturally occurring systems, such as in power, economics [10], [11], neural networks [16], and circuit theory [14]; in the last instance, they may additionally be used to model hysteresis [15]. They also provide a convenient form for the dynamical equations of large-scale interconnected systems [21]. Even the usual state and costate equations for optimal control are singular if the control weighting matrix is singular [1]. For a survey of linear singular systems, see [9].

Although, under certain conditions, some semistate variables may be eliminated to reduce the system to the usual state-space formulation, there are several good reasons for not doing so. Among these is the loss of sparsity, both in the system matrices and of the physical meaning of the variables. Some cases where the state equations do not exist may also be useful from a physical point of view.

Moreover, the state-space formulation has some notorious deficiencies. For example, these systems are closed under neither system inversion, where derivatives of the output are generally required, nor derivative feedback, which classical control theory has shown to be useful in many cases.

* Received by the editors February 17, 1987; accepted for publication (in revised form) August 30, 1988. This research was supported by National Science Foundation grant ECS-8518164.

† Department of Electrical and Electronic Engineering, Bogazici University, 80815 Bebek, Istanbul, Turkey.

‡ School of Electrical Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0250.

It is, therefore, desirable to extend the state-space techniques of design, which have been so successful in the control of dynamical systems, to the singular case.

Unfortunately, in so doing we find a major problem. Even if a singular system is well defined (i.e., $\det(sE - A \not\equiv 0)$ so that there exist unique solutions $x(t)$ for all $x(0)$ and suitable $u(t)$), it may be ill defined after the application of feedback of the proportional $(sE-(A+BF))$ or derivative $(s(E-BK)-A)$ sort. Therefore, the extension of state-space results is notoriously difficult.

In [23], a group-theoretic approach was used to extend the state-space geometric theory to singular systems, providing solutions to many design problems for these systems. However, because the feedback used was of a constrained form (e.g., "constant-ratio proportional-plus-derivative" (CRPD)), all of the limitations of state-space systems, such as the two mentioned above, were also extended to singular systems.

Our goal in writing this paper was to provide a rigorous foundation for a geometric theory for singular systems under proportional feedback (since such a feedback is generally easier to implement than proportional-plus-derivative feedback). This case is more difficult to deal with than the CRPD case. To achieve our goal, we rigorously define some basic geometric entities for these systems, exploring their properties, and examining some important distinctions that do not occur in the state-space case.

For simplicity in the following, we assume a regular (i.e., well-defined) system. Then, for completeness, at the end of the paper we extend all results to the nonregular case.

Let $\mathbf{X}$ and $\mathbf{U}$ be finite-dimensional vector spaces and let $A:\mathbf{X}\to\mathbf{X}$ and $B:\mathbf{U}\to\mathbf{X}$ be linear maps. A subspace $\mathbf{R}\subset\mathbf{X}$ is said to be a *reachability subspace* of the pair (A, B) if there exist two linear maps $F:\mathbf{X}\to\mathbf{U}$ and $G:\mathbf{U}\to\mathbf{U}$ such that $\mathbf{R}$ is the reachable subspace of the linear, time-invariant system defined by $(A+BF, BG)$ [28].

The notion of reachability subspaces of (A, B) has been demonstrated to be a very applicable as well as versatile tool in analysis and synthesis for linear systems, and consequently, it remains one of the major achievements of linear system theory. For a thorough discussion of the reachability subspaces and of the practical problems to which the concept has been successfully applied, the reader is referred to the classical text by Wonham [28].

One of the major results of the geometric system theory answers the question: Exactly when is a given subspace $\mathbf{R}\subset\mathbf{X}$ a reachability subspace of (A, B)?

THEOREM 1.1 [28]. *Let* $\mathbf{B}$ *denote the image of* B. $\mathbf{R}\subset\mathbf{X}$ *is a reachability subspace of* (A, B) *if and only if*:

(1) $A\mathbf{R}\subset\mathbf{R}+\mathbf{B}$;

(2) $\mathbf{R}=\lim_k \mathbf{S}_k$, *where* $\mathbf{S}_k$ *is defined by* $\mathbf{S}_{k+1}=\mathbf{R}\cap\{A\mathbf{S}_k+\mathbf{B}\}$; $\mathbf{S}_0=0$.

It is the aim of this paper to show that the pioneering work of Wonham and Morse on reachability subspaces can be extended to encompass the linear, time-invariant singular system

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

where $x(t)\in\mathbf{X}$, $u(t)\in\mathbf{U}$, and E, A, and B are linear maps with $\det(E)$ possibly equal to zero. For a discussion of the diversified disciplines where singular systems do arise naturally, see [1], [9]. If the constant-ratio proportional-plus-derivative feedback $u=F(x\cos\theta-\dot{x}\sin\theta)$ (with θ a parameter) is allowed, the extension of geometric notions to singular systems is straightforward [23]. However, our aim is to use only proportional feedback (i.e., $K=0$). In this case, as we shall see, the situation is far more challenging.

Now, let α be the index of nilpotency of E and let $C^{\alpha-1}$ and $C_p^{\alpha-1}$ denote the spaces of $(\alpha-1)$ times continuously differentiable and $(\alpha-1)$ times piecewise continuously differentiable mappings of $\mathbf{R}$ into $\mathbf{U}$, with $\mathbf{R}$ the real numbers. We consider both spaces as embedded subspaces of D', the space of distributions on $\mathbf{R}$ (with range space in $\mathbf{X}$). We also let D'_p denote the space of piecewise continuous distributions (again, with range space in $\mathbf{X}$).

Equation (1.1), or simply the pair (E, A), is said to be *regular* [5] if $\det(\lambda E - A) \not\equiv 0$. It will be our standing assumption throughout the paper that (1.1) is a regular system. Given any initial condition $x(0) = x_0$ and any $u \in C_p^{\alpha-1}$, there exists a unique distributional solution $x \in D'_p$ of (1.1) if and only if (1.1) is regular. Then, $x \in D'_p$ is the solution of (1.1) in the following sense. If x_+ and u_+ denote the (unique) restrictions of x and u to $[0, \infty)$ and if $x[\tau]$ denotes the impulsive part of x at τ (as defined in [3]), then $E\dot{x}_+ = Ax_+ + Bu_+ + \delta E x_0$, $x(0) = x_0$, and $E\dot{x}[\tau] = Ax[\tau] - \delta E \Delta_\tau x$, where δ is the Dirac delta and $\Delta_\tau x$ is the jump in x at τ. For imporous definitions of these entities and for details, see [3] and also [2], [24].

Here we would like to point out that this formulation renders both of the feedback laws $u = Fx + v$ and $u = K\dot{x} + Fx + v$, $v \in C_p^{\alpha-1}$ as admissible inputs.

Noting that it is really Ex_0 rather than x_0 that determines x_+, we say that a point $y \in X$ is *reachable* from $Ex_0 \in EX$ if there exists $u \in C^{\alpha-1}$ and $T > 0$ such that the solution $x(t)$ does not contain an impulsive part on $[0, T)$, is continuously differentiable on $(0, T)$, and satisfies $x(T) = y$. It is shown in [17] that there exists a subspace, called the *reachable subspace* of (1.1) and denoted by $\langle E, A|\mathbf{B}\rangle$, such that $y \in \mathbf{X}$ can be reached from the origin if and only if $y \in \langle E, A|\mathbf{B}\rangle$. Similarly, the origin can be reached from an Ex_0 if and only if $Ex_0 \in E\langle E, A|\mathbf{B}\rangle$ [17]. Consequently, $E\langle E, A|\mathbf{B}\rangle$ is called the *controllable subspace* of (1.1). Equation (1.1) is called *reachable* if $\langle E, A|\mathbf{B}\rangle = \mathbf{X}$ and controllable if $E\langle E, A|\mathbf{B}\rangle = E\mathbf{X}$ (or, equivalently, if $\langle E, A|\mathbf{B}\rangle + \mathbf{Ker}\ \mathbf{E} = \mathbf{X}$). Reachability (respectively, controllability) is equivalent to the absence of both finite and infinite zeros of the pencil $[\lambda E - AB]$ in the sense of Rosenbrock (respectively, Verghese) [3], [8].

The question that will be posed and investigated in the sections to follow is: Given a subspace $R \subset \mathbf{X}$, how and when can two linear maps $F: \mathbf{X} \to \mathbf{U}$ and $G: \mathbf{U} \to \mathbf{U}$ be found so that (1) $(E, A + BF)$ is regular; and (2) $\mathbf{R}$ is the reachable subspace of $(E, A + BF, BG)$? Note that a careless choice of F may yield a closed-loop system that is not regular. As the definitions of reachability and controllability depend on regularity, we need to guarantee the regularity of the closed-loop system to be able to talk about its reachable subspace. If there exists F and G to yield $\mathbf{R} = \langle E, A + BF|\mathbf{BG}\rangle$ with $(E, A + BF)$ regular, then $\mathbf{R}$ will be denoted a *reachability subspace of* (E, A, B) or a *generalized reachability subspace.*

Naturally, we will also define a subspace $C \subset E\mathbf{X}$ to be a *controllability subspace of* (E, A, B) or a *generalized controllability subspace* if $\mathbf{C} = E\mathbf{R}$ for some generalized reachability subspace $\mathbf{R}$. It should be clear that once a characterization of generalized reachability subspaces is available, a similar characterization for generalized controllability subspaces follows immediately. Therefore, we will mainly be interested in a geometric characterization of generalized reachability subspaces.

Notation. In this paper, vector spaces and their subspaces are written in boldface capital letters. Vectors are denoted by lowercase and linear maps are denoted by uppercase letters. If $E: \mathbf{X} \to \mathbf{Y}$ is a linear map, then the image of E is written as $\mathbf{E}$. If $\mathbf{S} \subset \mathbf{Y}$, then $E^{-1}\mathbf{S}$ is the inverse image of E (i.e., $\{x \in X : Ex \in S\}$), and $\mathbf{Ker}\ \mathbf{E} = E^{-1}\mathbf{0}$. The restriction of E to a subspace $\mathbf{R}$ is written as $E|\mathbf{R}$, and $E|\mathbf{R}$ with restricted codomain $\mathbf{S}$ is denoted by $\mathbf{S}|E|\mathbf{R}$. If $\mathbf{R} \subset \mathbf{X}$, then $\mathbf{R}_\mathbf{E} := \mathbf{R} \cap \mathbf{Ker}\ \mathbf{E}$. If $\mathbf{S}$ is a vector space, then $d\mathbf{S}$

denotes its dimension. The symbol + stands for direct sum of subspaces (or of linear maps). If q is a positive integer, then $\underline{q}$ denotes the set $\{1, 2, \cdots, q\}$. If $\{S_K\}$ is a sequence of subspaces and if $\mathbf{S}$ is a subspace, then $\{\mathbf{S_K}\}\uparrow\mathbf{S}$ (respectively, $\{\mathbf{S_K}\}\downarrow\mathbf{S}$) means that $\{\mathbf{S_K}\}$ is nondecreasing (respectively, nonincreasing) and converges to $\mathbf{S}$. Where the operations of subspace intersection ($\cap$) and summation ($+$) occur without parentheses in a single expression (e.g., $\mathbf{RS}+\mathbf{T}$), we use the convention that the intersection is performed first. Finally, the reader should note that the shorthand notation A_F will sometimes be used for $A+BF$.

2. Background and problem formulation. Let $\mathbf{K}$ and $\mathbf{Q}$ be fixed but otherwise arbitrary subspaces of $\mathbf{X}$. A subspace $\mathbf{R} \subset \mathbf{K}$ is said to be an $(A, E, \mathbf{Q})$-*invariant subspace of* $\mathbf{K}$ if $\mathbf{AR} \subset E\mathbf{R}+\mathbf{Q}$. If $\mathbf{R}$ satisfies $\mathbf{R} = \mathbf{K} \cap E^{-1}(A\mathbf{R}+\mathbf{Q})$, then $\mathbf{R}$ is called a *restricted* $(E, A, \mathbf{Q})$-*invariant* (or, in short, an $(E, A, \mathbf{Q})_r$-inv.) *subspace of* $\mathbf{K}$. The class of all $(A, E, \mathbf{Q})$-inv. subspaces of $\mathbf{K}$ contains a unique supremal element $\mathbf{V}^*(A, E, \mathbf{Q}; \mathbf{K})$ [19]. The class of all $(E, A, \mathbf{Q})_r$-inv. subspaces of $\mathbf{K}$ contains a unique least member $\mathbf{S}_*(E, A, \mathbf{Q}; \mathbf{K})$ [19]. If $\mathbf{V_k}$ and $\mathbf{S_k}$ are defined by

$$\mathbf{V}_{k+1} = \mathbf{K} \cap A^{-1}(E\mathbf{V}_k + \mathbf{Q}), \qquad \mathbf{V}_0 = \mathbf{K}, \tag{2.1}$$

$$\mathbf{S}_{k+1} = \boldsymbol{K} \cap E^{-1}(A\mathbf{S}_k + \mathbf{Q}), \qquad \mathbf{S}_0 = \mathbf{K}\ \mathbf{Ker}\ \mathbf{E}, \tag{2.2}$$

then $\{\mathbf{V}_k\}\downarrow\mathbf{V}^*(A, E, \mathbf{Q}; \mathbf{K})$ and $\{\mathbf{S_k}\}\uparrow S_*(E, A, \mathbf{Q}; \mathbf{K})$ [19].

In the sequel, (2.1) and (2.2) will be performed with different maps and $\mathbf{Q}$'s. To economize in space, the following abbreviations will be used:

$$\mathbf{V}_0^* := \mathbf{V}^*(A, E, \mathbf{0}; \mathbf{X}), \qquad \mathbf{S}_*^0 := \mathbf{S}_*(E, A, \mathbf{0}; \mathbf{X}),$$

$$\mathbf{V}_{F,0}^* := \mathbf{V}^*(A+BF, E, \mathbf{0}; \mathbf{X}), \qquad \mathbf{S}_*^{F,0} := \mathbf{S}_*(E, A+BF, \mathbf{0}; \mathbf{X}),$$

$$\mathbf{V}^* := \mathbf{V}^*(A, E, \mathbf{B}; \mathbf{X}), \qquad \mathbf{S}_* := \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{X}),$$

$$\mathbf{V}_{F,G}^* := \mathbf{V}^*(A+BF, E, \mathbf{BG}; \mathbf{X}), \qquad \mathbf{S}_*^{F,G} := \mathbf{S}_*(E, A+BF, \mathbf{BG}; \mathbf{X}).$$

$\mathbf{V}_0^*$ and $\mathbf{S}_*^0$ were first studied in [27]. If $\sigma(E, A)$ denotes $\{\lambda \in C;\ \det(\lambda E - A) = 0\}$, then $\mathbf{V}_0^*$ is the direct sum of the eigenspaces corresponding to $\lambda \in \sigma(E, A)$ [17] and is called the *initial manifold* of (E, A) [27]; $\mathbf{S}_*^0$ on the other hand, is called the *final manifold* of (E, A) [27], and is the eigenspace corresponding to the unbounded eigenvalue of E and A [17]. The import of the analysis in [27] is the following theorem that presents what seems to be the only geometric test for the regularity of (E, A).

THEOREM 2.1 [27].

(1) (E, A) *is regular if and only if* $\mathbf{Ker}\ \mathbf{E} \cap \mathbf{V}_0^* = \mathbf{0}$;

(2) *If* (E, A) *is regular, then* $\mathbf{V}_0^* \oplus \mathbf{S}_*^0 = \mathbf{X}$;

(3) $\sigma(E, A) = \sigma(E|\mathbf{V}_0^*, A|\mathbf{V}_0^*)$ *and* $\mathrm{d}\mathbf{V}_0^* = \deg|\lambda E - A|$.

$\mathbf{V}^*$ and $\mathbf{S}_*$ have been defined in [19], where they have been shown to be instrumental in a coordinate-free characterization of the reachable subspace $\langle E, A|\mathbf{B}\rangle$ of (E, A, B). Generalizing Schumacher's algebraic characterization of almost invariance [22], Malabre has shown in [12] that if $Q = B$, then (2.2) generates the *supremal (generalized) almost reachability subspace contained in* $\mathbf{K}$. Thus, a subspace $\mathbf{R}$ that satisfies $\mathbf{R} = \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{R})$ will be said to be a (*generalized*) *almost reachability subspace.* The reader should also note the initial condition of the recursion (2.2), which was taken to be $\mathbf{0}$ in [19], has been modified in light of the analysis given in [12].

This change, however, does not affect the limit of the recursion. The main result of [19] is the following.

THEOREM 2.2 [19].

(1) $\langle E, A|\mathbf{B}\rangle = \mathbf{V}^* \cap \mathbf{S}_*$;

(2) $\langle E, A|\mathbf{B}\rangle = \mathbf{V}^*(A, E, \mathbf{B}; \mathbf{S}_*)$.

The following result was stated as a conjecture in [19].

THEOREM 2.3. $\langle E, A|\mathbf{B}\rangle = \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{V}^*)$.

Proof. See the Appendix for the proof.

An immediate consequence of Theorem 2.3 that will prove to be crucial in our discussion is given by the following corollary.

COROLLARY 2.1. $\langle E, A|\mathbf{B}\rangle = \mathbf{S}_*(E, A, \mathbf{B}; \langle E, A|\mathbf{B}\rangle)$. *That is, if* $\mathbf{S}_k$ *is defined by*

$$\mathbf{S}_{k+1} = \langle E, A|\mathbf{B}\rangle \cap E^{-1}\{A\mathbf{S}_k + \mathbf{B}\}, \qquad \mathbf{S}_0 = \langle E, A|\mathbf{B}\rangle \cap \mathbf{Ker}\ \mathbf{E}, \tag{2.3}$$

then $\{\mathbf{S}_k\} \uparrow \langle E, A|\mathbf{B}\rangle$.

Proof. Let $\mathbf{S}_k$ be defined by (2.3) and define $\tilde{S}_k$ by

$$\tilde{\mathbf{S}}_{k+1} = \mathbf{V}^* \cap E^{-1}\{A\tilde{S}_k + \mathbf{B}\}, \qquad \tilde{\mathbf{S}}_0 = \mathbf{V}^* \cap \mathbf{Ker}\ \mathbf{E},$$

then, by Theorem 2.3, $\{\tilde{\mathbf{S}}_k\} \uparrow \langle E, A|\mathbf{B}\rangle$. Hence, $\tilde{\mathbf{S}}_k \subset \langle E, A|\mathbf{B}\rangle$ for all k, and therefore

$$\tilde{\mathbf{S}}_{k+1} = \langle E, A|\mathbf{B}\rangle \cap \mathbf{V}^* \cap E^{-1}\{A\tilde{S}_k + \mathbf{B}\}. \tag{2.4}$$

Using Theorem 2.2(1) in (2.4), we have

$$\tilde{\mathbf{S}}_{k+1} = \langle E, A|\mathbf{B}\rangle E^{-1}\{A\tilde{\mathbf{S}}_k + \mathbf{B}\}.$$

Then, $\mathbf{S}_k = \tilde{\mathbf{S}}_k$ for all k. Consequently, $\lim_k \mathbf{S}_k = \lim_k \tilde{\mathbf{S}}_k$. That is, $\{\mathbf{S}_k\} \uparrow \langle E, A|\mathbf{B}\rangle$. □

In view of Theorems 2.1 and 2.2(1), the question of whether or not a given subspace $R \subset \mathbf{X}$ is a generalized reachability subspace can now be formulated as follows.

Problem formulation. Given $\mathbf{R} \subset \mathbf{X}$, find, if and when possible, two linear maps $F: \mathbf{X} \to \mathbf{U}$ and $G: U \to U$ so that if $\mathbf{V}^*_{F,0}$, $\mathbf{V}^*_{F,G}$, and $\mathbf{S}^{F,G}_*$ denote $\lim_k \mathbf{V}^k_{F,0}$, $\lim_k \mathbf{V}^k_{F,G}$, and $\lim_k \mathbf{S}^{F,G}_k$, respectively, where

$$\mathbf{V}^{k+1}_{F,0} = (A + BF)^{-1}E\mathbf{V}^k_{F,0}, \qquad \mathbf{V}^0_{F,0} = \mathbf{X}, \tag{2.5}$$

$$\mathbf{V}^{k+1}_{F,G} = (A + BF)^{-1}\{E\mathbf{V}^k_{F,G} + \mathbf{BG}\}, \qquad \mathbf{V}^0_{F,G} = \mathbf{X}, \tag{2.6}$$

$$\mathbf{S}^{F,G}_{k+1} = E^{-1}\{(A + BF)\mathbf{S}^{F,G}_k + \mathbf{BG}\}, \qquad \mathbf{S}^{F,G}_0 = \mathbf{Ker}\ \mathbf{E}, \tag{2.7}$$

then we have the following:

(1) $\mathbf{Ker}\ \mathbf{E} \cap \mathbf{V}^*_{F,0} = \mathbf{0}$;

(2) $\mathbf{R} = \mathbf{V}^*_{\mathbf{F},\mathbf{G}} \cap \mathbf{S}^{F,G}_*$.

The solution of this problem will involve a generalization of the notion of the *friends* of an (A, B)-invariant subspace [28] that will be discussed in § 3.

It is clear from the problem formulation that what is at hand is a set of two intertwined problems formulated by (1) and (2) above. The problem of choosing F to render $(E, A + BF)$ regular will be tackled in § 3. The latter problem, as given by (2), will be solved in § 4.

Before closing this section, we present three properties that will find very frequent applications in the sequel. Let $\mathbf{R}$, $\mathbf{S}$, and $\mathbf{Q}$ be subspaces of $\mathbf{X}$.

PROPERTY 2.1. *If* $\mathbf{R} \subset \mathbf{S}$, *then* $\mathbf{S} \cap (\mathbf{R} + \mathbf{Q}) = \mathbf{R} + \mathbf{S} \cap \mathbf{Q}$. *This is* (3.1.b) of [28, p. 4].

PROPERTY 2.2. $E^{-1}(\mathbf{R} + \mathbf{S}) = E^{-1}\mathbf{R} + E^{-1}\mathbf{S}$ *if and only if* $\mathbf{E} \cap (\mathbf{R} + \mathbf{S}) = \mathbf{E} \cap \mathbf{R} + \mathbf{R} \cap \mathbf{S}$. (See [28, p. 8].)

PROPERTY 2.3. $\mathbf{R} \cap E^{-1}(A\mathbf{S} \cap E\mathbf{R} + \mathbf{Q}) = \mathbf{R} \cap E^{-1}A\mathbf{S} + \mathbf{R} \cap E^{-1}\mathbf{Q}$.

Proof. By Property 2.2,

$$\begin{aligned}\mathbf{R}\cap E^{-1}(A\mathbf{S}\cap E\mathbf{R}+\mathbf{Q}) &= \mathbf{R}\cap\{E^{-1}(A\mathbf{S}\cap E\mathbf{R})+E^{-1}\mathbf{Q}\}\\ &= \mathbf{R}\cap\{E^{-1}A\mathbf{S}\cap E^{-1}E\mathbf{R}+E^{-1}\mathbf{Q}\}\\ &= \mathbf{R}\cap\{E^{-1}A\mathbf{S}\cap(\boldsymbol{R}+\mathbf{Ker\ E})+E^{-1}\mathbf{Q}\}.\end{aligned}$$

$\mathbf{Ker\ E}\subset E^{-1}A\mathbf{S}$ and therefore, by Property 2.1,

$$\begin{aligned}\mathbf{R}\cap E^{-1}(A\mathbf{S}\cap E\mathbf{R}+\mathbf{Q}) &= \mathbf{R}\cap\{E^{-1}A\mathbf{S}\cap\mathbf{R}+\mathbf{Ker\ E}+E^{-1}\mathbf{Q}\}\\ &= \mathbf{R}\cap\{E^{-1}A\mathbf{S}\cap\mathbf{R}+E^{-1}\mathbf{Q}\}.\end{aligned}$$

Applying Property 2.1 one more time,

$$= \mathbf{R}\cap E^{-1}A\mathbf{S}+\mathbf{R}\cap E^{-1}\mathbf{Q}. \qquad \square$$

3. Friends and regular friends of an (A, E, B)-invariant subspace. We start our discussion by recalling that a linear map is said to be a "friend" of an $(A, \mathbf{B})$-inv. subspace $\mathbf{V}$ if $(A+BF)\mathbf{V}\subset\mathbf{V}$ [28]. A naive generalization of the notion of "friends" of an $(A, \mathbf{B})$-inv. subspace [28] motivates the definition of a linear map $F:\mathbf{X}\to\mathbf{U}$ as a friend of a given subspace $R\subset\mathbf{X}$ if $(A+BF)\mathbf{R}\subset E\mathbf{R}$. Let $\mathbf{F}(\mathbf{R})$ denote the class of all friends of $\mathbf{R}$. It follows easily that $\mathbf{F}(\mathbf{R})\neq\phi$ if and only if $\mathbf{R}$ is $(A, E, \mathbf{B})$-inv.

The necessity of choosing F to make $(E, A+BF)$ regular, and the following proposition, which states that some friends of $\mathbf{R}$ may be "bad friends," severely limit the applicability of the notion of "friends" to singular systems.

PROPOSITION 3.1. *Let* $\mathbf{R}$ *be* $(A, E, \mathbf{B})$*-invariant. If* $\mathbf{R}\cap\mathbf{Ker\ E}\neq\mathbf{0}$ *and if* $F\in\mathbf{F}(\mathbf{R})$, *then* $(E, A+BF)$ *is not regular.*

Proof. Define $\mathbf{R}_E := \mathbf{R}\cap\mathbf{Ker\ E}$ and assume $\mathbf{R}_E\neq\mathbf{0}$. If $F\in\mathbf{F}(\mathbf{R})$, then $(A+BF)\mathbf{R}$ $E\mathbf{R}$; that is, $\mathbf{R}$ is $(A+BF, E, \mathbf{0})$-inv. As $\mathbf{V}^*_{F,0}$ is the supremal $(A+BF, E, \mathbf{0})$-inv. subspace of $\mathbf{X}$, R $\mathbf{V}^*_{F,0}$. Therefore, $\mathbf{Ker\ E}\cap\mathbf{V}^*_{F,0}\supset\mathbf{R}_E\neq\mathbf{0}$ and, by Theorem 2.1(1), $(E, A+BF)$ is not regular. $\square$

We remark that in general we cannot expect that $\mathbf{R}\cap\mathbf{Ker\ E}=\mathbf{0}$. To remedy the situation, the following definition is introduced. It will become clear in the sequel that the central role played by friends of an $(A, \mathbf{B})$-inv. subspace in the analysis of reachability subspaces of (A, B) will, in the context of singular systems, be played by *regular friends* of an $(A, E, \mathbf{B})$-inv. subspace, as next defined.

DEFINITION 3.1. A linear map $F:\mathbf{X}\to\mathbf{U}$ is called a *regular friend* of a subspace $\mathbf{R}\subset\mathbf{X}$ if (1) $(E, A+BF)$ is regular; and (2) F is a friend of $\mathbf{R}_1$ for some $\mathbf{R}_1$ satisfying $\mathbf{R}=\mathbf{R}_1\oplus\mathbf{R}_E$.

The class of all regular friends of $\mathbf{R}$ will be denoted by $\mathbf{RF}(\mathbf{R})$. Clearly, if $\mathbf{R}_E=\mathbf{0}$, then $\mathbf{RF}(\mathbf{R})=\mathbf{F}(\mathbf{R})$. If $\mathbf{R}_E\neq\mathbf{0}$, then Proposition 3.1 yields $\mathbf{RF}(\mathbf{R})\cap\mathbf{F}(\mathbf{R})=\varnothing$. However, it is not because $\mathbf{RF}(\mathbf{R})=\varnothing$ when $\mathbf{R}_E\neq 0$ that the intersection is empty. Indeed, by explicitly constructing a regular friend, the following theorem establishes that $\mathbf{RF}(\mathbf{R})\neq\varnothing$ whenever R is (A, E, B)-inv.

THEOREM 3.1. *If* R *is* $(A, E, \mathbf{B})$*-invariant, then* $\mathbf{RF}(\mathbf{R})\neq\varnothing$.

Proof. Recall that $\mathbf{V}^*$ is the supremal $(A, E, \mathbf{B})$-inv. subspace of $\mathbf{X}$. Then $\mathbf{R}\subset\mathbf{V}^*$. Write $\mathbf{R}=\mathbf{R}_1\oplus\mathbf{R}_E$ for some $\mathbf{R}_1$. As $\mathbf{R}_1\cap\mathbf{Ker\ E}=\mathbf{0}$, there exists a $\mathbf{V}$ satisfying (1) $\mathbf{R}_1\subset\mathbf{V}$; and (2) $\mathbf{V}^*=\mathbf{V}\oplus\mathbf{V}^*\cap\mathbf{Ker\ E}$. Let $\mathbf{V}^*_E := \mathbf{V}^*\cap\mathbf{Ker\ E}$. Note that $E\mathbf{R}=E\mathbf{R}_1$ and $E\mathbf{V}^*=E\mathbf{V}$. Then $A\mathbf{R}_1\subset E\mathbf{R}_1+\mathbf{B}$ and $A\mathbf{V}\subset E\mathbf{V}+\mathbf{B}$. Choose $K_1:\mathbf{X}\to\mathbf{U}$ to satisfy $(A+BK_1)\mathbf{V}\subset\mathbf{EV}$. Then $(A+BK_1)\mathbf{R}_1\subset A\mathbf{R}_1+\mathbf{B}\subset E\mathbf{R}_1+\mathbf{B}$ and let a $K_2:\mathbf{R}_1\to\mathbf{U}$ be chosen to yield $(A+BK_1+BK_2)\mathbf{R}_1\subset E\mathbf{R}_1$. Let $K_2\equiv 0$ on a complement of R_1 in X. Then it immediately follows that $K_1+K_2\in\mathbf{F}(\mathbf{R}_1)\cap\mathbf{F}(\mathbf{V})$. Let $F_0 := K_1+K_2$.

It follows from [19] that $\mathbf{V}^*$ satisfies $d(E\mathbf{V}^*+\mathbf{B}) = d\mathbf{V}^*$. Write $E\mathbf{V}^*+\mathbf{B} = E\mathbf{V}^* \oplus \tilde{\mathbf{B}}$ for some $\tilde{\mathbf{B}} \subset \mathbf{B}$. Then $d\tilde{\mathbf{B}} = d(E\mathbf{V}^*+\mathbf{B}) - dE\mathbf{V}^* = d\mathbf{V}^* - dE\mathbf{V}^* = d\mathbf{V}_E^*$. Define $q := d\mathbf{V}_E^*$. Let $\{v_i: i \in \mathbf{q}\}$ and $\{B\omega_i: i \in \mathbf{q};\ \omega_i \in \mathbf{U}\}$ be bases for $\mathbf{V}_E^*$ and $\tilde{\mathbf{B}}$, respectively.

Note that $(A+BF_0)\ \mathbf{V}_E^* \subset A\mathbf{V}_E^* + \mathbf{B} \subset A\mathbf{V}^* + \mathbf{B} \subset E\mathbf{V}^* + \mathbf{B} = E\mathbf{V} + \mathbf{B}$. Then there exists $\bar{v}_i \in \mathbf{V}$ and $u_i \in \mathbf{U}$ such that

$$(A+BF_0)v_i = E\bar{v}_i + Bu_i, \qquad i \in \mathbf{q}.$$

Define $F_1: \mathbf{V}_E^* \to \mathbf{U}$ by

$$F_1 v_i = -u_i + \omega_i, \qquad i \in \mathbf{q}.$$

Note that $(A+BF_0+BF_1)\mathbf{V}_E^* \subset E\mathbf{V}^* + \tilde{\mathbf{B}}$, but $(A+BF_0+BF_1)\mathbf{V}_E^* \cap E\mathbf{V}^* = \mathbf{0}$. Extend F_1 to $\mathbf{X}$ by defining $F_1 \equiv 0$ on a complement of $\mathbf{V}_E^*$ in $\mathbf{X}$. Finally, let $F \equiv F_0 + F_1$. Then $(A+BF)\mathbf{R}_1 = (A+BF_0)\mathbf{R}_1 \subset E\mathbf{R}_1$ because $F_0 \in \mathbf{F}(\mathbf{R}_1)$. Therefore $F \in \mathbf{F}(\mathbf{R}_1)$, too. Also note that $(A+BF)\mathbf{V} = (A+BF_0)\mathbf{V} \subset E\mathbf{V}$, and consequently, $F \in \mathbf{F}(\mathbf{V})$. It remains to show that $(E, A+BF)$ is regular.

Recall that $\mathbf{V}_{F,0}^*$ denotes the supremal $(A+BF, E, \mathbf{0})$-inv. subspace of $\mathbf{X}$. As $\mathbf{V}_{F,0}^*$ satisfies $(A+BF)\mathbf{V}_{F,0}^* \subset E\mathbf{V}_{F,0}^*$, there follows $A\mathbf{V}_{F,0}^* \subset E\mathbf{V}_{F,0}^* + \mathbf{B}$, i.e., $\mathbf{V}_{F,0}^* \subset \mathbf{V}^*$. Then $(A+BF)\mathbf{V}_{F,0}^* \subset E\mathbf{V}_{F,0}^* E\mathbf{V}^*$. This and $(A+BF)\mathbf{V}_E^* \cap E\mathbf{V}^* = 0$ imply $\mathbf{V}_{F,0}^* \cap \mathbf{V}_E^* = 0$. Then, as $\mathbf{V}_{F,0}^* \subset \mathbf{V}^*$, $\mathbf{V}_{F,0}^* \cap \mathbf{Ker}\ \mathbf{E} = \mathbf{V}_{F,0}^* \cap \mathbf{V}^* \cap \mathbf{Ker}\ \mathbf{E} = \mathbf{V}_{F,0}^* \cap \mathbf{V}_E^* = 0$. Therefore, by Theorem 2.1(1), $(E, A+BF)$ is regular. Thus, $F \in \mathbf{RF}(\mathbf{R})$. □

Remarks. (1) The reader should note that the last paragraph of the proof above shows that if F satisfies $A_F V_E^* \cap E\mathbf{V}^* = \mathbf{0}$ and $dA_F\mathbf{V}_E^* = d\mathbf{V}_E^*$, then (E, A_F) is regular.

(2) It is by construction that F is a regular friend of $\mathbf{V}^*$ also. Thus, if R is $(A, E, \mathbf{B})$-inv., then $\mathbf{RF}(\mathbf{V}^*) \cap \mathbf{RF}(\mathbf{R}) \neq 0$.

COROLLARY 3.1. *Define* $\mathbf{V}^*(\mathbf{R}) := \mathbf{V}^*(A, E, B; \mathbf{R})$. *Then* $\mathbf{RF}(\mathbf{R}) \neq 0$ *if and only if* $\mathbf{R} = \mathbf{V}^*(\mathbf{R}) + \mathbf{R} \cap \mathbf{Ker}\ \mathbf{E}$.

Proof. If $F \in \mathbf{RF}(\mathbf{R})$, then $A_F\mathbf{R}_1 \subset E\mathbf{R}_1$ for some $\mathbf{R}_1$ satisfying $\mathbf{R} = \mathbf{R}_1 \oplus \mathbf{R} \cap \mathbf{Ker}\ \mathbf{E}$. As $A_F\mathbf{R}_1 \subset E\mathbf{R}_1$ implies $A\mathbf{R}_1 \subset E\mathbf{R}_1 + \mathbf{B}$, we have $\mathbf{R}_1 \subset \mathbf{V}^*(\mathbf{R})$. Consequently, $R = \mathbf{V}^*(\mathbf{R}) + \mathbf{R} \cap \mathbf{Ker}\ \mathbf{E}$. To prove the converse, simply note that, by Theorem 3.1, $\mathbf{RF}(\mathbf{V}^*(\mathbf{R})) \neq 0$ and by definition, any regular friend of $\mathbf{V}^*(\mathbf{R})$ is also a regular friend of $\mathbf{R}$. □

Some of the properties of regular friends of $\mathbf{R}$ are presented below.

PROPOSITION 3.2. *If* $F \in R\mathbf{F}(\mathbf{R})$ *and if* $A_F := A+BF$, *then the following are true*:

(1) $\mathbf{R}_1 \subset \mathbf{V}_{F,0}^*$ *for some* $\mathbf{R}_1$ *satisfying* $\mathbf{R} = \mathbf{R}_1 \oplus \mathbf{R}_E$;

(2) $d(A_F\mathbf{R}_E) = d\mathbf{R}_E$;

(3) $A_F\mathbf{R}_E \cap E\mathbf{R} = \mathbf{0}$;

(4) $A_F\mathbf{R} + E\mathbf{R} = A_F\mathbf{R}_E + E\mathbf{R} \subset \mathbf{R} + \mathbf{B}$;

(5) $d(A_F\mathbf{R} + E\mathbf{R}) = d\mathbf{R}$;

(6) $A_F\mathbf{R} + E\mathbf{R} + \hat{\mathbf{B}}$ *for some* $\hat{\mathbf{B}} \subset B$ *with* $d\hat{\mathbf{B}} = d\mathbf{R}_E$;

(7) $\sigma(E|\mathbf{R}, A_F|\mathbf{R}) = \sigma(E|\mathbf{V}^*(\mathbf{R}), A_F|\mathbf{V}^*(\mathbf{R})) = \sigma(E|\mathbf{R}_1, A_F|\mathbf{R}_1)$;

(8) *If* $\mathbf{K} \in \mathbf{RF}(\mathbf{R})$ *also, then* $B(K-F)\mathbf{R} \subset B \cap (E\mathbf{R} + A_F\mathbf{R})$.

Proof. (1) The proof of (1) is immediate from Definition 3.1 and the fact that $\mathbf{V}_{F,0}^*$ is the supremal $(A_F, E, \mathbf{0})$-inv. subspace of $\mathbf{X}$.

(2) Note that $\mathbf{V}_{F,0}^*$ satisfies $\mathbf{V}_{F,0}^* = A_F^{-1}E\mathbf{V}_{F,0}^*$. Then $\mathbf{Ker}\ \mathbf{A_F} \subset \mathbf{V}_{F,0}^*$. If $F \in \mathbf{RF}(\mathbf{R})$, then $\mathbf{Ker}\ \mathbf{E} \cap \mathbf{Ker}\ \mathbf{A_F} \subset \mathbf{Ker}\ \mathbf{E} \cap \mathbf{V}_{F,0}^* = \mathbf{0}$ (by Theorem 2.1(1)). Therefore, $d(A_F\mathbf{R}_E) = d\mathbf{R}_E$.

(3) Recall that $\mathbf{S}_*^{F,0}$, the least $(E, A_F, \mathbf{0})_r$-inv. subspace of $\mathbf{X}$, satisfies $\mathbf{S}_*^{F,0} = E^{-1}A_F\mathbf{S}_*^{F,0}$. Now, let $\mathbf{R}_1$ be as in (1) and let $x \in A_F\mathbf{R}_E \cap E\mathbf{R}$. Then $x = A_F r_E = Er_1$ for some $r_E \in \mathbf{R}_E$ and $r_1 \in \mathbf{R}_1$. Then $r_1 \in E^{-1}A_F R_E \subset E^{-1}A_F\mathbf{Ker}\ \mathbf{E} \subset E^{-1}A_F\mathbf{S}_*^{F,0} = \mathbf{S}_*^{F,0}$, and therefore $r_1 \in \mathbf{R}_1 \cap \mathbf{S}_*^{F,0}$. Since (E, A_F) is regular, Theorem 2.1(2) implies $r_1 = 0$, which in turn implies $x = 0$.

(4) Note that if $\mathbf{R}_1$ is as in (1), then $A_F\mathbf{R}_1 \subset E\mathbf{R}_1$ and $A_F\mathbf{R} + E\mathbf{R} = A_F\mathbf{R}_E + E\mathbf{R}_1$. Also, $A_F\mathbf{R} + E\mathbf{R} \subset E\mathbf{R} + \mathbf{B} + E\mathbf{R} = E\mathbf{R} + \mathbf{B}$.

(5) $d(A_F\mathbf{R} + E\mathbf{R}) = d(A_F\mathbf{R}_E + E\mathbf{R}) = d(A_F\mathbf{R}_E) + dE\mathbf{R} - d(A_F\mathbf{R}_E \cap E\mathbf{R})$. Note that $E\mathbf{R} = E\mathbf{R}_1$. As $\mathbf{R}_1 \subset \mathbf{V}^*_{F,0}$ and as $\mathbf{Ker}\,\mathbf{E} \cap \mathbf{V}^*_{F,0} = \mathbf{0}$, $d(E\mathbf{R}) = d\mathbf{R}_1$. Then the result follows from (2) and (3).

(6) Write $(A_F\mathbf{R} + E\mathbf{R}) \cap B = E\mathbf{R} \cap \mathbf{B} \oplus \hat{\mathbf{B}}$ for some $\hat{\mathbf{B}} \subset \mathbf{B}$. Note that (4) implies

$$A_F R + E\mathbf{R} = (A_F R + E\mathbf{R}) \cap (E\mathbf{R} + \mathbf{B}).$$

Using Property 2.1, we have

$$\begin{aligned} A_F\mathbf{R} + E\mathbf{R} &= E\mathbf{R} + (A_F\mathbf{R} + E\mathbf{R}) \cap B \\ &= E\mathbf{R} + E\mathbf{R} \cap \mathbf{B} \oplus \hat{\mathbf{B}} \\ &= E\mathbf{R} \oplus \hat{\mathbf{B}}. \end{aligned}$$

Then, (5) implies $d\mathbf{R} = d(E\mathbf{R} + \hat{\mathbf{B}})$. As $\hat{\mathbf{B}}\ E\mathbf{R} = \mathbf{0}$, $d\hat{\mathbf{B}} = d\mathbf{R} - dE\mathbf{R} = d\mathbf{R}_E$.

(7) The first equality follows immediately from Theorem 2.1(3) and the fact that the initial manifold of $(E|\mathbf{R}, A_F|\mathbf{R})$ is contained in $\mathbf{V}^*(\mathbf{R})$. To prove the second equality, write $\mathbf{R} = \mathbf{R}_1 \oplus \mathbf{R}_E$ and let v_i be the eigenvector corresponding to $\lambda_i \in \sigma(E|\mathbf{R}, A_F|\mathbf{R})$. Writing $v_i = r_i + r^i_E$, where $r_i \in \mathbf{R}_1$ and $r^i_E \in \mathbf{R}_E$, we have $\lambda_i E r_i = A_F r_i + A_F r^i_E$. As $A_F r_i \in A_F\mathbf{R}_1 \subset E\mathbf{R}_1$, we have, by (3) above, $r^i_E = 0$. Thus, $v_i \in \mathbf{R}_1$ and, therefore, $\sigma(E|R, A_F|R) = \sigma(E|\mathbf{R}_1, A_F|\mathbf{R}_1)$.

(8) Let $\mathbf{K} \in \mathbf{RF}(\mathbf{R})$. Then by (5),

$$\begin{aligned} d\mathbf{R} &= d[E\mathbf{R} + (A + BK)\mathbf{R}] \\ &= d[E\mathbf{R} + (A + BF)\mathbf{R} + B(K - F)\mathbf{R}]. \end{aligned}$$

Since $F \in \mathbf{RF}(\mathbf{R})$, $d\mathbf{R} = d[E\mathbf{R} + (A + BF)\mathbf{R}]$. Thus, $B(K - F)\mathbf{R} \subset \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})$. □

4. Generalized reachability subspaces. Recall that a subspace $R \subset \mathbf{X}$ was defined in § 1 to be a generalized reachability subspace if there existed linear maps $F: \mathbf{X} \to \mathbf{U}$ and $G: U \to U$ such that (1) $(E, A + BF)$ was regular; and (2) $\mathbf{R}$ was the reachable subspace of $(E, A + BF, BG)$.

An immediate consequence of the definition and Corollary 2.1 is the following theorem.

THEOREM 4.1. *If R is a generalized reachability subspace, then $\mathbf{R}$ is $(A, E, \mathbf{B})$-invariant and $\mathbf{R} = \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{R})$.*

Proof. If there exist F and G such that $\mathbf{R}$ is the reachable space of $(E, A + BF, BG)$, then, by Theorem 2.2(2), $\mathbf{R}$ satisfies $\mathbf{R} = \mathbf{S}^{F,G}_* \cap (A + BF)^{-1}\{E\mathbf{R} + \mathbf{BG}\}$. Then $(A + BF)\mathbf{R} \subset E\mathbf{R} + \mathbf{BG}$ implying $AR \subset E\mathbf{R} + \mathbf{B}$.

Corollary 2.1 shows that $R = \mathbf{S}_*(E, A + BF, \mathbf{BG}; \mathbf{R})$. Note that $\mathbf{R} = \mathbf{S}_*(E, A + BF, \mathbf{BG}; \mathbf{R}) \subset \mathbf{S}_*(E, A + BF, \mathbf{B}; \mathbf{R})$ because $\mathbf{BG} \subset \mathbf{B}$. On the other hand, it is by definition that $\mathbf{S}_*(E, A + BF, \mathbf{B}; \mathbf{R}) \subset \mathbf{R}$. Therefore, $\mathbf{R} = \mathbf{S}_*(E, A + BF, \mathbf{BG}; \mathbf{R}) \subset \mathbf{S}_*(E, A + BF, \mathbf{B}; \mathbf{R}) \subset \mathbf{R}$. Then $\mathbf{R} = \mathbf{S}_*(E, A + BF, \mathbf{B}; \mathbf{R})$. As $(A + BF)\mathbf{S}_k + \mathbf{B}$ for any $\mathbf{S}_k \subset \boldsymbol{X}$, it trivially follows that $S_*(E, A, \mathbf{B}; \mathbf{R}) = \mathbf{S}_*(E, A + BF, \mathbf{B}; \mathbf{R}) = \mathbf{R}$. Hence the theorem is proved. □

The reader should remember that $\mathbf{R} = \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{R})$ means $\mathbf{R} = \lim_k \mathbf{S}_k$, where $\mathbf{S}_k$ is defined by

$$\mathbf{S}_{k+1} = \mathbf{R} \cap E^{-1}\{A\mathbf{S}_k + \mathbf{B}\}, \qquad \mathbf{S}_0 = \mathbf{R} \cap \mathbf{Ker}\,\mathbf{E}. \tag{4.1}$$

It is not only because the algorithm (4.1) reduces to the well-known almost-reachability subspace algorithm of Willems [29] in case $E = I$, but also because of the reasoning given in [12] that we call (4.1) the generalized almost reachability subspace algorithm (GARSA). A subspace $\mathbf{R}$ will be said to satisfy GARSA if $\lim_k \mathbf{S}_k = \mathbf{R}$; that is to say,

$\mathbf{R}$ is the *supremal generalized almost reachability subspace contained* in $\mathbf{R}$. By proving that the converse of Theorem 4.1 is also true, Theorem 4.2 below will show that $\mathbf{R}$ is a generalized reachability subspace if and only if $\mathbf{R}$ is an $(A, E, \mathbf{B})$-inv. generalized almost-reachability subspace. However, before proceeding to Theorem 4.2, we need to prove two propositions.

PROPOSITION 4.1. *Let an* $(A, E, \mathbf{B})$*-invariant subspace* $\mathbf{R}$ *satisfy* GARSA. *Given an* $F \in \mathbf{RF}(\mathbf{R})$, *choose* $G: \mathbf{U} \to \mathbf{U}$ *so that* $\mathbf{BG} = \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})$. *Then* $\mathbf{R} = \mathbf{S}_*(E, A+BF, \mathbf{BG}; \mathbf{R})$.

Proof. Let $F \in \mathbf{RF}(\mathbf{R})$. Then $F \in \mathbf{F}(\mathbf{R}_1)$ for some $\mathbf{R}_1$ satisfying $\mathbf{R} = \mathbf{R}_1 \oplus \mathbf{R}_E$. (Recall that $\mathbf{R}_E = \mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}$.) Note that $\mathbf{S}_*(E, A, \mathbf{B}; \mathbf{R}) = \mathbf{S}_*(E, A+BF, \mathbf{B}; \mathbf{R}) = \mathbf{R}$. Let $\mathbf{B}^{(1)} = \mathbf{B}$, $\mathbf{B}^{(2)} = (E\mathbf{R} + A_F\mathbf{R}) \cap B$, and define $\mathbf{S}_k^{(i)}$ for $i = 1, 2$ by

$$\mathbf{S}_{k+1}^{(i)} = \mathbf{R} \cap E^{-1}\{A_F\mathbf{S}_k^{(i)} + \mathbf{B}^{(i)}\}, \qquad \mathbf{S}_0^{(i)} = \mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}$$

where, as before, A_F denotes $(A+BF)$. Clearly, if $\mathbf{S}_*^{(1)}(\mathbf{R}) := \lim_k \mathbf{S}_k^{(1)}$ and $\mathbf{S}_*^{(2)}(\mathbf{R}) := \lim_k \mathbf{S}_k^{(2)}$, then $\mathbf{S}_*^{(2)}(\mathbf{R}) \subset \mathbf{S}_*^{(1)}(\mathbf{R})$ because $\mathbf{B}^{(2)} \subset \mathbf{B}^{(1)}$. On the other hand, $\mathbf{S}_0^{(1)} \subset \mathbf{S}_*^{(2)}(\mathbf{R})$ and if $\mathbf{S}_k^{(1)} \subset \mathbf{S}_*^{(2)}(\mathbf{R})$, then $\mathbf{S}_{k+1}^{(1)} \subset \mathbf{R} \cap E^{-1}\{A_F\mathbf{S}_*^{(2)}(\mathbf{R}) + \mathbf{B}\}$. Let $x \in \mathbf{S}_{k+1}^{(1)}$. Then, $x \in \mathbf{R}$ and $Ex = A_F s + Bu$ for some $u \in \mathbf{U}$ and $s \in \mathbf{S}_*^{(2)}(\mathbf{R})$. Then, $Bu = Ex - A_F s \in E\mathbf{R} + A_F\mathbf{S}_*^{(2)}(\mathbf{R}) \subset E\mathbf{R} + A_F\mathbf{R}$ and, therefore, we have $x \in \mathbf{R} \cap E^{-1}\{A_F\mathbf{S}_*^{(2)}(\mathbf{R}) + \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})\}$. Since $x \in \mathbf{S}_{k+1}^{(1)}$ is arbitrary, we conclude that $\mathbf{S}_{k+1}^{(1)} \subset \mathbf{R} \cap E^{-1}\{A_F\mathbf{S}_*^{(2)}(\mathbf{R}) + \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})\} = \mathbf{S}_*^{(2)}(\mathbf{R})$. This proves that $\mathbf{S}_k^{(1)} \subset \mathbf{S}_*^{(2)}(\mathbf{R})$ for all k. Thus $\mathbf{S}_*^{(1)}(\mathbf{R}) \subset \mathbf{S}_*^{(2)}$. This and $\mathbf{S}_*^{(2)}(\mathbf{R}) \subset \mathbf{S}_*^{(1)}(\mathbf{R})$ together imply $\mathbf{S}_*^{(1)}(\mathbf{R}) = \mathbf{S}_*^{(2)}(\mathbf{R})$. Then, defining G by $\mathbf{BG} = (E\mathbf{R} + A_F\mathbf{R}) \cap B$ completes the proof. □

PROPOSITION 4.2. *Let an* $(A, E, \mathbf{B})$*-invariant subspace* $\mathbf{R}$ *satisfy* GARSA. *Let* $F \in \mathbf{RF}(\mathbf{R})$. *Write* $R = \mathbf{R}_1 \oplus \mathbf{R}_E$, *where* $F \in \mathbf{F}(\mathbf{R}_1)$ *and* $\mathbf{R}_E = \mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}$. *If* G *is chosen to satisfy* $\mathbf{BG} = (E\mathbf{R} + A_F\mathbf{R}) \cap \mathbf{B}$, *then*:

(1) $\mathbf{S}_*^{F,G} = \mathbf{S}_*^{F,0} + \mathbf{R}_1$;

(2) $\mathbf{V}_{F,G}^* = \mathbf{V}_{F,0}^* + \mathbf{R}_E$.

Remark. Recall that if $\mathbf{V}_{F,G}^k$, $\mathbf{V}_{F,0}^0$, $\mathbf{S}_k^{F,G}$, and $\mathbf{S}_k^{F,0}$ are defined by

$$\mathbf{V}_{F,G}^{k+1} = A_F^{-1}\{E\mathbf{V}_{F,G}^k + \mathbf{BG}\}, \qquad \mathbf{V}_{F,G}^0 = \mathbf{X}, \tag{4.2}$$

$$\mathbf{V}_{F,0}^{k+1} = A_F^{-1}\{E\mathbf{V}_{F,0}^k\}, \qquad \mathbf{V}_{F,0}^0 = \mathbf{X}, \tag{4.3}$$

$$\mathbf{S}_{k+1}^{F,G} = E^{-1}\{A_F\mathbf{S}_k^{F,G} + \mathbf{BG}\}, \qquad \mathbf{S}_0^{F,G} = \mathbf{Ker}\, \mathbf{E}, \tag{4.4}$$

$$\mathbf{S}_{k+1}^{F,0} = E^{-1}A_F\mathbf{S}_k^{F,0}, \qquad \mathbf{S}_0^{F,0} = \mathbf{Ker}\, \mathbf{E}, \tag{4.5}$$

then $\{\mathbf{V}_{F,G}^k\} \downarrow \mathbf{V}_{F,G}^*$, $\{\mathbf{V}_{F,0}^k\} \downarrow \mathbf{V}_{F,0}^*$, $\{\mathbf{S}_k^{F,G}\} \uparrow \mathbf{S}_*^{F,G}$, and $\{\mathbf{S}_k^{F,0}\} \uparrow \mathbf{S}_*^{F,0}$. It should be clear that $\mathbf{V}_{F,0}^* \subset \mathbf{V}_{F,G}^*$ and $\mathbf{S}_*^{F,0} \subset \mathbf{S}_*^{F,G}$.

Proof. (1) Note that $\mathbf{S}_*(E, A_F, \mathbf{BG}; \mathbf{R}) \subset \mathbf{S}_*(E, A_F, \mathbf{BG}; \mathbf{X}) = \mathbf{S}_*^{F,G}$. Since R satisfies GARSA, $\mathbf{R} = \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{R})$, and by Proposition 4.1, $\mathbf{R} = \mathbf{S}_*(E, A_F, \mathbf{BG}; \mathbf{R})$. Therefore

$$\mathbf{R} = \mathbf{S}_*(E, A_F, \mathbf{BG}; \mathbf{R}) \subset \mathbf{S}_*(E, A_F, \mathbf{BG}; \mathbf{X}) = \mathbf{S}_*^{F,G}.$$

This and $\mathbf{S}_*^{F,0} \subset \mathbf{S}_*^{F,G}$ imply $\mathbf{S}_*^{F,0} + \mathbf{R} \subset \mathbf{S}_*^{F,G}$. To prove $\mathbf{S}_*^{F,G} \subset \mathbf{S}_*^{F,0} + \mathbf{R}$ define $\mathbf{S}_k$ by

$$\begin{aligned} \mathbf{S}_{k+1} &= E^{-1}\{A_F\mathbf{S}_k + \mathbf{BG}\} \\ &= E^{-1}\{A_F\mathbf{S}_k + \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})\}, \\ \mathbf{S}_0 &= \mathbf{Ker}\, \mathbf{E}, \end{aligned}$$

and note that $\mathbf{S}_0 \subset \mathbf{S}_*^{F,0} + \mathbf{R}$. If $\mathbf{S}_k \subset \mathbf{S}_*^{F,0} + \mathbf{R}$ is assumed, then

$$\begin{aligned} \mathbf{S}_{k+1} &\subset E^{-1}\{A_F\mathbf{S}_*^{F,0} + A_F\mathbf{R} + \mathbf{B} \cap (E\mathbf{R} + A_F\mathbf{R})\} \\ &\subset E^{-1}\{A_F\mathbf{S}_*^{F,0} + A_F\mathbf{R} + E\mathbf{R}\} \\ &= E^{-1}\{A_F\mathbf{S}_*^{F,0} + A_F\mathbf{R}_E + E\mathbf{R}\} \quad \text{(by Proposition 3.2(4))} \\ &= E^{-1}\{A_F\mathbf{S}_*^{F,0} + E\mathbf{R}\}. \end{aligned}$$

The last equality follows from the fact that $R_E = \mathbf{R} \cap \mathbf{Ker}\, \mathbf{E} \subset E^{-1}A_F\mathbf{S}_*^{F,0} = \mathbf{S}_*^{F,0}$. Then, by Property 2.3, we have

$$\begin{aligned}\mathbf{S}_{k+1} &= E^{-1}A_F\mathbf{S}_*^{F,0} + \mathbf{R} + \mathbf{Ker}\, \mathbf{E} \\ &= E^{-1}A_F\mathbf{S}_*^{F,0} + \mathbf{R} \\ &= \mathbf{S}_*^{F,0} + \mathbf{R}.\end{aligned}$$

Thus $\mathbf{S}_*^{F,G} = \lim_k \mathbf{S}_k \subset \mathbf{S}_*^{F,0} + \mathbf{R}$. Together with $\mathbf{S}_*^{F,0} + \mathbf{R} \subset \mathbf{S}_*^{F,G}$, this implies $\mathbf{S}_*^{F,G} = \mathbf{S}_*^{F,0} + \mathbf{R}$. Then $\mathbf{S}_*^{F,G} = \mathbf{S}_*^{F,0} + \mathbf{R}_1 + \mathbf{R}_E = \mathbf{S}_*^{F,0} + \mathbf{R}_1$.

(2) Let $\mathbf{V}_{F,G}^k$ be defined by (4.2). As $\mathbf{V}_{F,0}^* \subset \mathbf{V}_{F,G}^*$ and $\{\mathbf{V}_{F,G}^k\}$ is nonincreasing, there follows $\mathbf{V}_{F,0}^* \subset \mathbf{V}_{F,G}^* \subset \mathbf{V}_{F,G}^k$ for all k. Let $\mathbf{R}_1$ be as before. As $A_F\mathbf{R}_1 \subset E\mathbf{R}_1$, $\mathbf{R}_1 \subset \mathbf{V}_{F,0}^*$ because $\mathbf{V}_{F,0}^*$ is the supremal $(A_F, E, \mathbf{0})$-inv. subspace of $\mathbf{X}$. Then $E\mathbf{R} = E\mathbf{R}_1 \subset E\mathbf{V}_{F,0}^* \subset E\mathbf{V}_{F,G}^* \subset E\mathbf{V}_{F,G}^k$ for all k. Thus (4.2) can be rewritten as

$$\mathbf{V}_{F,G}^{k+1} = A_F^{-1}\{E\mathbf{V}_{F,G}^k + E\mathbf{R} + (A_F\mathbf{R} + E\mathbf{R}) \cap B\}$$

and the proofs of Proposition 3.2(4) and (6) imply

$$\begin{aligned}\mathbf{V}_{F,G}^{k+1} &= A_F^{-1}\{E\mathbf{V}_{F,G}^k + E\mathbf{R} + A_F\mathbf{R}_E\} \\ &= A_F^{-1}\{E\mathbf{V}_{F,G}^k + A_F\mathbf{R}_E\}.\end{aligned}$$

Property 2.3, applied to the equations above, yields

$$\begin{aligned}\mathbf{V}_{F,G}^{k+1} &= A_F^{-1}E\mathbf{V}_{F,G}^k + \mathbf{R}_E + \mathbf{Ker}\, \mathbf{A_F} \\ &= A_F^{-1}E\mathbf{V}_{F,G}^k + \mathbf{R}_E;\end{aligned} \tag{4.6}$$

$\mathbf{R}_E \subset \mathbf{V}_{F,G}^k$ for all k yields $\mathbf{R}_E \subset \mathbf{V}_{F,G}^*$. This and $\mathbf{V}_{F,0}^* \subset \mathbf{V}_{F,G}^*$ together show that $\mathbf{V}_{F,0}^* + \mathbf{R}_E \subset \mathbf{V}_{F,G}^*$. If $\mathbf{V}_{F,0}^k$ is defined by (4.3), then $\mathbf{V}_{F,G}^0 \subset \mathbf{V}_{F,0}^0 + \mathbf{R}_E$. Assume $\mathbf{V}_{F,G}^k \subset \mathbf{V}_{F,0}^k + \mathbf{R}_E$. Then by (4.6),

$$\begin{aligned}\mathbf{V}_{F,G}^{k+1} &\subset A_F^{-1}E(\mathbf{V}_{F,0}^k + \mathbf{R}_E) + \mathbf{R}_E \\ &= A_F^{-1}E\mathbf{V}_{F,0}^k + \mathbf{R}_E \\ &= \mathbf{V}_{F,0}^{k+1} + \mathbf{R}_E.\end{aligned}$$

As $\mathbf{V}_{F,G}^k \subset \mathbf{V}_{F,0}^k + \mathbf{R}_E$ for all k, there follows $\mathbf{V}_{F,G}^* \mathbf{V}_{F,0}^* + \mathbf{R}_E$. This, together with $\mathbf{V}_{F,G}^* \supset \mathbf{V}_{F,0}^* + \mathbf{R}_E$, proves (2). □

The main result of this paper can now be proved.

THEOREM 4.2. *A subspace* $\mathbf{R} \subset \mathbf{X}$ *is a generalized reachability subspace if and only if*:

(1) $A\mathbf{R} \subset E\mathbf{R} + \mathbf{B}$;

(2) $\mathbf{R} = \lim_k \mathbf{S}_k$, *where*

$$\mathbf{S}_{k+1} = \mathbf{R} \cap E^{-1}\{A\mathbf{S}_k + \mathbf{B}\}, \qquad \mathbf{S}_0 = \mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}.$$

Proof. Necessity was proved in Theorem 4.1. To prove sufficiency, let $F \in \mathbf{RF}(\mathbf{R})$ and define G by $BG = (A_F\mathbf{R} + E\mathbf{R}) \cap B$. By Theorem 2.2(1) $\langle E, A + BF | \mathbf{BG} \rangle$, the reachable subspace of $(E, A + BF, BF)$, is given by $\mathbf{S}_*^{F,G} \cap \mathbf{V}_{F,G}^*$. Proposition 4.2 implies

$$\langle E, A + BF | \mathbf{BG} \rangle = (\mathbf{S}_*^{F,0} + \mathbf{R}_1) \cap (\mathbf{V}_{F,0}^* + \mathbf{R}_E) \tag{4.7}$$

where $R_1 + \mathbf{R}_E = \mathbf{R}$ and $(A + BF)\mathbf{R}_1 \subset E\mathbf{R}_1$, i.e., $\mathbf{R}_1 \subset \mathbf{V}_{F,0}^*$. Then $\mathbf{R}_1 \subset \mathbf{V}_{F,0}^* + \mathbf{R}_E$ and Property 2.1, applied to (4.7), yields

$$\langle E, A + BF | \mathbf{BG} \rangle = \mathbf{S}_*^{F,0} \cap (\mathbf{V}_{F,0}^* + \mathbf{R}_E) + \mathbf{R}_1.$$

As $\mathbf{R}_E$ $\mathbf{Ker}\, \mathbf{E}$ $\mathbf{S}_*^{F,0}$ (see (4.5)), it follows by Property 2.1 that

$$\langle E, A + BF | \mathbf{BG} \rangle = \mathbf{R}_E + \mathbf{S}_*^{F,0} \cap \mathbf{V}_{F,0}^* + \mathbf{R}_1.$$

Because $F \in \mathbf{RF}(\mathbf{R})$, $(E, A + BF)$ is regular. Then, by Theorem 2.1(2), $\mathbf{S}_*^{F,0} \cap \mathbf{V}_{F,0}^* = \mathbf{0}$ and the result follows. □

Now let $\mathbf{K} \subset \mathbf{X}$ be a subspace and let $\mathbf{R}(E, A, \mathbf{B}; \mathbf{K})$ denote the class of all generalized reachability subspaces contained in $\mathbf{K}$. As $\mathbf{0}$ is a reachability subspace, $\mathbf{R}(E, A, \mathbf{B}; \mathbf{K})$ is nonempty, and the proof of Lemma 5.5 in [28] with only very minor modifications shows that $\mathbf{R}(E, A, \mathbf{B}; \mathbf{K})$ is closed under the operation of subspace addition. Therefore, $\mathbf{K}$ contains a unique supremal generalized reachability subspace that will be denoted by $\mathbf{R}^*(\mathbf{K})$.

Recall that $\mathbf{V}^*(A, E, \mathbf{B}; \mathbf{K})$ denotes supremal $(A, E, \mathbf{B})$-inv. subspace contained in $\mathbf{K}$ and can be computed by the recursion

$$\mathbf{V}_{k+1} = \mathbf{K} \cap A^{-1}\{E\mathbf{V}_k + \mathbf{B}\}, \qquad \mathbf{V}_0 = \mathbf{K}.$$

Let $\mathbf{V}^*(\mathbf{K}) := \mathbf{V}^*(A, E, \mathbf{B}; \mathbf{K})$. Then the proof of the following theorem mimics the proof of Theorem 5.6 in [28] and will be left to the reader.

THEOREM 4.3. $\mathbf{R}^*(\mathbf{K})$ *is given by* $\lim_k \mathbf{R}_k$, *where*

$$\mathbf{R}_{k+1} = \mathbf{V}^*(\mathbf{K}) \cap E^{-1}\{A\mathbf{R}_k + \mathbf{B}\} \qquad \mathbf{R}_0 = \mathbf{V}^*(\mathbf{K}) \cap \mathbf{Ker}\ \mathbf{E}.$$

Recall that in § 1, a subspace $\mathbf{C}$ was defined to be a generalized controllability subspace if $\mathbf{C} = E\mathbf{R}$ for some generalized reachability subspace $\mathbf{R}$. More precisely we have Theorem 4.4.

THEOREM 4.4. $\mathbf{C}$ *is a generalized controllability subspace if and only if* $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{C})$.

Proof. If $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{C})$, then $\mathbf{C}$ is clearly a generalized controllability subspace. If $\mathbf{C} = E\mathbf{R}$ for some generalized reachability subspace $\mathbf{R} \subset E^{-1}\mathbf{C}$, then

$$\mathbf{C} = E\mathbf{R} \subset E\mathbf{R}^*(E^{-1}\mathbf{C}) \subset EE^{-1}\mathbf{C} \subset \mathbf{C}$$

and consequently, $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{C})$. □

The class of all generalized controllability subspaces contained in a given subspace $\mathbf{K}$ contains $\mathbf{0}$ and is closed under the operation of subspace addition. Thus, it contains a unique supremal element $\mathbf{C}^*(\mathbf{K})$. Theorem 4.5 follows immediately from Theorem 4.4.

THEOREM 4.5. $\mathbf{C}^*(\mathbf{K}) = E\mathbf{R}^*(E^{-1}\mathbf{K})$.

5. On spectral assignability. The equivalence of reachability and controllability for (1.1) to $E = I$ is well known. Therefore in this case it is unnecessary to distinguish between reachability and controllability subspaces. Indeed, the accepted practice is to talk about the property of controllability and controllability subspaces of (A, B) [28]. However, this simplification in the terminology of proper state-space systems has the adverse effect of concealing the relations between different characterizations of controllability subspaces and the different properties that produce such characterizations. Indeed, the dynamic characterization [28] of a controllability subspace $\mathbf{R}$ of (I, A, B) as having the property that any $x \in \mathbf{R}$ can be reached from the origin in finite time along a smooth trajectory, generated by a smooth input and not leaving $\mathbf{R}$, clearly and explicitly reflects the property of reachability. In light of our discussion in the previous sections, it should be clear that a similar dynamic characterization applies to reachability subspaces, though not necessarily to controllability subspaces, of (E, A, B).

On the other hand, a very important result [28] states that $\mathbf{R}$ with $d\mathbf{R} = q$ is a controllability subspace of (I, A, B) if and only if, for every symmetric set Λ of q complex numbers, there exists a friend F of $\mathbf{R}$ such that the spectrum $\alpha((A + RF)|R)$ of $(A + BF)|\mathbf{R}$, the restriction of $(A + BF)$ to $\mathbf{R}$, is precisely Λ. In this section, we show that this characterization depends on the property of controllability (as opposed to reachability) by showing that an extension of the characterization applies to controllability subspaces but not, in general, to reachability subspaces of (E, A, B). Indeed, the following theorem, which has already appeared in the literature [4], [7], [13] explicitly

or implicitly, justifies this comment immediately. However, our proof, which is geometric in nature, is new.

THEOREM 5.1. *Given any symmetric set* Λ *of* $r := \text{rank } E$ *complex numbers, there exists a linear map* $F: X \to U$ *so that* $\Lambda = \sigma(E, A_F)$ *if and only if* (E, A, B) *is controllable.*

Remarks. (1) Note that $\Lambda = \sigma(E, A_F)$ implies that $\sigma(E, A_F) \neq C$, that is, (E, A_F) is regular.

(2) Note that reachability of (E, A, B) - the condition $\langle E, A|\mathbf{B}\rangle = \mathbf{X}$ - implies controllability, that is, the condition $E\langle E, A|\mathbf{B}\rangle = E\mathbf{X}$, but not vice versa. Thus, reachability of (E, A, B) is sufficient but not necessary for the arbitrary assignment of $\sigma(E, A_F)$.

Proof. Let Λ have r distinct elements and assume that $\Lambda \cap \sigma(E, A) = \varnothing$. Let F assign Λ as $\sigma(E, A_F)$. For $\lambda_i \in \Lambda$, let v_i be the corresponding eigenvector. Then $v_i \in \langle E, A|B\rangle$ and $\{Ev_i: i \in \underline{r}\}$ is a linearly independent set [18]. Then $E\langle E, A|\mathbf{B}\rangle = E\mathbf{X}$ and (E, A, B) is controllable.

Let (E, A, B) be controllable. Then $\langle E, A|\mathbf{B}\rangle + \mathbf{Ker}\ \mathbf{E} = \mathbf{X}$ and thus, $V^* + \mathbf{Ker}\ \mathbf{E} = \mathbf{X}$. Let F_0 be a regular friend of $\mathbf{V}^*$. Then F_0 assigns some $\mathbf{V}$ satisfying $\mathbf{V} \oplus \mathbf{V}^* \cap \mathbf{Ker}\ \mathbf{E} = \mathbf{V}^*$ as the initial manifold of a regular system (E, A_{F_0}, B). Then, by Theorem 2.1, $\mathbf{V} \oplus \mathbf{Ker}\ \mathbf{E} = \mathbf{X}$ and $E\mathbf{V} \oplus A_{F_0}\mathbf{Ker}\ \mathbf{E} = \mathbf{X}$. Let V and N be basis matrices for $\mathbf{V}$ and $\mathbf{Ker}\ \mathbf{E}$. Let $P = [VN]$ and $Q = [EVA_FN]$. Then $(Q^{-1}EP, Q^{-1}A_FP, Q^{-1}B)$ becomes

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} A_1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u \tag{5.1}$$

where $x_1 \in \mathbf{R}^r$ and $x_2 \in \mathbf{R}^{n-r}$.

As controllability is preserved by regular friends and by coordinate transformations, (5.1) is controllable, implying that (A_1, B_1) is a controllable pair [2], [3]. Then, there exists an $F_1: \mathbf{R}^r \to \mathbf{U}$ such that $\Lambda = \sigma(A_1 + B_1F_1)$. Let $F_2: \mathbf{X} \to \mathbf{U}$ be the extension of F_1, which is zero on $\mathbf{R}^{n-r}$. Then

$$\det \begin{bmatrix} \lambda I - A_1 - B_1F_1 & 0 \\ -B_2F_1 & I \end{bmatrix} = \det (\lambda I - A_1 - B_1F_1).$$

Thus, $F = F_0 + F_2P^{-1}$ is the map that yields $\Lambda = \sigma(E, A_F)$. □

In light of Theorem 5.1, the following result is not unexpected.

THEOREM 5.2. *Let* $\mathbf{C} \subset E\mathbf{X}$ *be a subspace with* $d\mathbf{C} = q \geqq 1$. $\mathbf{C}$ *is a generalized controllability subspace if and only if, given any symmetric set* Λ *of* q *complex numbers, there exists a linear map* $F: \mathbf{X} \to \mathbf{U}$ *so that*:

(1) F *is a regular friend of* $E^{-1}\mathbf{C}$;

(2) $\Lambda = \sigma(E|E^{-1}\mathbf{C}, A_F|E^{-1}\mathbf{C})$.

Proof. Let $\mathbf{C}$ be a generalized controllability subspace, i.e., $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{C})$. Let $F_0 \in \mathbf{RF}(\mathbf{R}^*(E^{-1}\mathbf{C})) \cap \mathbf{RF}(\mathbf{V}^*)$ be as constructed in Theorem 3.1. Since $E^{-1}\mathbf{C} = \mathbf{R}^*(E^{-1}\mathbf{C}) + \mathbf{Ker}\ \mathbf{E}$, F_0 is also a regular friend of $E^{-1}\mathbf{C}$. Define $\bar{\mathbf{C}} = EE^{-1}\mathbf{C} + A_{F_0}E^{-1}\mathbf{C}$ and let $E_0 := \bar{C}|E|E^{-1}\mathbf{C}$, $A_0 := \bar{\mathbf{C}}|A_{F_0}|E^{-1}\mathbf{C}$ and choose G so that $B_0 = BG$ satisfies $\mathbf{B}_0 = \mathbf{B} \cap \bar{\mathbf{C}}$. Now

$$\begin{aligned} \mathbf{R}^*(E^{-1}\mathbf{C}) &= \mathbf{R}^*(E, A, B; E^{-1}\mathbf{C}) \\ &= \mathbf{R}^*(E, A_{F_0}, \mathbf{B}; E^{-1}\mathbf{C}) \\ &= \mathbf{R}^*(E, A_{F_0}, \mathbf{B}_0; E^{-1}\mathbf{C}) \quad \text{(Proposition 4.1)} \\ &= \mathbf{R}^*(E_0, A_0, \mathbf{B}_0; E^{-1}\mathbf{C}). \end{aligned}$$

Since $E_0\mathbf{R}^*(E_0, A_0, \mathbf{B}_0; E^{-1}\mathbf{C}) = E_0\mathbf{R}^*(E^{-1}\mathbf{C}) = E_0E^{-1}\mathbf{C} = \mathbf{E}_0$, then (E_0, A_0, B_0) is controllable. Then, by Theorem 5.1, given any symmetric set Λ of rank $E_0 = dE_0E^{-1}\mathbf{C} = q$ complex numbers, there exists an $F_1: E^{-1}\mathbf{C} \to \mathbf{U}$ such that $\Lambda = \sigma(E_0, A_0 + B_0F_1)$.

It remains to extend F_1 to $\mathbf{X}$ without destroying the regularity of the closed-loop system. To that end, let $\mathbf{R}$ denote the initial manifold of $(E_0, A_0+B_0F_1)$. By Theorem 2.1(3), $d\mathbf{R} = \deg|\Lambda E_0 - (A_0+B_0F_1)| = q$. This and regularity of $(E_0, A_0+B_0F_1)$ imply that $\mathbf{R}\oplus\mathbf{Ker\ E} = \mathbf{R}\oplus\mathbf{Ker\ E} = E^{-1}\mathbf{C}$. Let $F_2 = F_1$ on $\mathbf{R}$ and let $F_2 = 0$ on a subspace which complements $\mathbf{R}$ to $\mathbf{X}$. Finally, let $F = F_0 + GF_2$. Since $F|\mathbf{V}_E^* = F_0|\mathbf{V}_E^*$ and since, by construction, F_0 satisfies $A_{F_0}\mathbf{V}_E^*\ E\mathbf{V}^* = \mathbf{0}$ and $dA_{F_0}\mathbf{V}_E^* = d\mathbf{V}_E^*$, (E, A_F) is regular (see the remark following Theorem 3.1). Also note that $(A+BF)\mathbf{R} = (A+B(F_0+GF_1))\mathbf{R} = [(A+BF_0)+B_0F_1]\mathbf{R} = (A_0+B_0F_1)\mathbf{R} \subset E_0\mathbf{R} = E\mathbf{R}$. Thus, F is a regular friend of $E^{-1}\mathbf{C}$. Also, $\Lambda = \sigma(E_0, A_0+B_0F_1) = \sigma(E|E^{-1}C, A_F|E^{-1}C)$.

To prove the converse statement, note that the existence of an F in $\mathbf{RF}(E^{-1}\mathbf{C})$ implies, through Corollary 3.1, that $E^{-1}\mathbf{C} = \mathbf{V}^*(E^{-1}\mathbf{C}) + \mathbf{Ker\ E}$. Fix $F_0 \in RF(\mathbf{V}^*|E^{-1}\mathbf{C}) \cap \mathbf{RF}(\mathbf{V}^*)$ (use Theorem 3.1 to construct F_0). If F is another regular friend of $E^{-1}\mathbf{C}$, then, by Proposition 3.2, $B(F-F_0)E^{-1}\mathbf{C} \subset \mathbf{B}\cap(EE^{-1}\mathbf{C} + A_{F_0}E^{-1}\mathbf{C}) = \mathbf{B}\cap(C + A_{F_0}E^{-1}\mathbf{C}) = \mathbf{B}_0$. Defining F_1: $E^{-1}\mathbf{C} \to \mathbf{U}$ by $B_0F_1 = B(F-F_0)E^{-1}\mathbf{C}$, we conclude that given any symmetric set Λ of q complex numbers, there exists an F_1 such that $\sigma(E_0, A_0+B_0F_1) = \Lambda$. Then, by Theorem 5.1, (E_0, A_0, B_0) is controllable. That is to say, $E_0\mathbf{R}^*(E_0, A_0, \mathbf{B}_0; E^{-1}\mathbf{C}) = \mathbf{E}_0 = E_0E^{-1}\mathbf{C} = \mathbf{C}$. Then the result follows by noting that $\mathbf{R}^*(E_0, A_0, \mathbf{B}_0; E^{-1}\mathbf{C}) = \mathbf{R}^*(E, A, \mathbf{B}; E^{-1}\mathbf{C})$. Thus, $\mathbf{C} = E\mathbf{R}^*(E^{-1}\mathbf{C})$ and $\mathbf{C}$ is a generalized controllability subspace. □

COROLLARY 5.1. *Let* $\mathbf{R}$ *be a generalized reachability subspace with* $dE\mathbf{R} = q \geqq 1$. *Given any symmetric set* Λ *of* q *complex numbers, there exists a regular friend* F *of* R *that renders* $\Lambda = \sigma(E|\mathbf{R}, A_F|\mathbf{R})$.

Proof. Note that $\mathbf{ER}$ is a controllability subspace, and $\mathbf{RF}(\mathbf{R}) \cap \mathbf{RF}(\mathbf{E}^{-1}\mathbf{ER}) \neq 0$. Then choose $F_0 \in \mathbf{RF}(\mathbf{R}) \cap \mathbf{RF}(\mathbf{E}^{-1}\mathbf{ER}) \cap \mathbf{RF}(\mathbf{V}^*)$. Then the proof of Theorem 5.2 also proves the corollary because $\sigma(E|R, A_F|\mathbf{R}) = \sigma(E|\mathbf{E}^{-1}\mathbf{ER}, A_F|\mathbf{E}^{-1}\mathbf{ER}) = \sigma(E|\mathbf{R}+\mathbf{Ker\ E}, A_F|\mathbf{R}+\mathbf{Ker\ E})$. □

As the nonequivalence of reachability and controllability for singular systems is due to the singularity of E, it may be conjectured that if $\mathbf{R}\cap\mathbf{Ker\ E} = 0$, then the converse of Corollary 5.1 is also true. This indeed is the case.

LEMMA 5.1. *Let* $\mathbf{R}$ *with* $d\mathbf{R} = q \geqq 1$ *satisfy* $\mathbf{R}\ \mathbf{Ker\ E} = \mathbf{0}$. *Suppose that, given any symmetric set* Λ *of* q *complex numbers, there exists a regular friend* F *of* $\mathbf{R}$ *that yields* $\Lambda = \sigma(E|\mathbf{R}, A_F|\mathbf{R})$. *Then* $\mathbf{R}$ *is a generalized reachability subspace.*

Proof. If $\mathbf{R}$ has a regular friend, then by Corollary 3.1, $\mathbf{R} = \mathbf{V}^*(\mathbf{R}) + \mathbf{R}\cap\mathbf{Ker\ E}$. As $\mathbf{R}\cap\mathbf{Ker\ E} = \mathbf{0}$, we have $\mathbf{R} = \mathbf{V}^*(\mathbf{R})$, that is, $\mathbf{R}$ is $(A, E, \mathbf{B})$-inv. To show that $R^*(E, A, \mathbf{B}; \mathbf{R}) = \mathbf{R}$, fix $F_0 \in RF(\mathbf{R}) \cap \mathbf{RF}(\mathbf{V}^*)$ and let $E_0 = E\mathbf{R} + A_{F_0}\mathbf{R}|E|\mathbf{R}$; $A_0 = E\mathbf{R} + A_{F_0}\mathbf{R}|A|\mathbf{R}$ and $\mathbf{B}_0 = \mathbf{B}\cap(E\mathbf{R} + A_F\mathbf{R})$. Then the premise of the lemma is equivalent to asserting the existence of an F_1: $R \to U$ that yields $\sigma(E_0, A_0+B_0F_1) = \Lambda$ for any given symmetric set Λ of q complex numbers (see the proof of Theorem 5.2). As rank $E_0 = dE\mathbf{R} = d\mathbf{R} = q$, we conclude by Theorem 5.1 that (E_0, A_0, B_0) is controllable, that is, $\langle E_0, A_0|\mathbf{B}_0\rangle + \mathbf{Ker\ E}_0 = \mathbf{R}$. As $\mathbf{Ker\ E}_0 = \mathbf{Ker\ E}\cap\mathbf{R} = \mathbf{0}$, we have $\langle E_0, A_0|\mathbf{B}_0\rangle = \mathbf{R}$. Then $\mathbf{S}_*(E_0, A_0, \mathbf{B}_0; \mathbf{R}) = \mathbf{R}$ and as $\mathbf{S}_*(E_0, A_0, \mathbf{B}_0; \mathbf{R}) = S_*(E, A, \mathbf{B}; \mathbf{R})$, $\mathbf{R}$ is a generalized almost-reachability subspace also. Thus R is a generalized reachability subspace. □

We shall end our discussion by investigating the effects of using proportional-plus-derivative feedback (rather than only proportional feedback) on the spectral assignability properties of generalized reachability subspaces. We first note Lemma 5.2.

LEMMA 5.2. $\mathbf{R}$ *is a generalized reachability subspace if and only if there exist* F, K, *and* G *such that* $\mathbf{R} = \langle E_k, A_F|\mathbf{BG}\rangle$.

Proof. (If.) Take $K = 0$ and F, G as in § 4. (Only if.) $\mathbf{R}$ is $(A_F, E_k, \mathbf{BG})$-inv. and, therefore, it is $(A, E, \mathbf{B})$-inv. Also, $\mathbf{S}_*(E_k, A_F, \mathbf{BG}; \mathbf{R}) = \mathbf{R}$ clearly implies $\mathbf{S}_*(E, A, \mathbf{R}; \mathbf{R}) = \mathbf{R}$. □

LEMMA 5.3. *There exists a linear map* $K: \mathbf{X} \to \mathbf{U}$ *such that if* R *is* $(A, E, \mathbf{B})$*-invariant, then* $\mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}_k = \mathbf{0}$.

Proof. Note that $d(E\mathbf{V}^* + \mathbf{B}) = d\mathbf{V}^*$[3], [20]. Write $E\mathbf{V}^* + \mathbf{B} = E\mathbf{V}^* \oplus \mathbf{B}_1$. Then $d(\mathbf{V}^* \cap \mathbf{Ker}\, \mathbf{E}) = d\mathbf{B}_1$. Let $p = d\mathbf{B}_1$ and choose $\{v_i: i \in \mathbf{p}\}$ and $\{Bw_i: i \in \mathbf{p}\}$ as bases for $\mathbf{V}^* \cap \mathbf{Ker}\, \mathbf{E}$ and B_1, respectively. Define $K_0 v_i = -w_i$, $i \in \mathbf{p}$ and let K be any extension of K_0 to $\mathbf{X}$. Then $E_k\mathbf{V}^* = E\mathbf{V}^* + \mathbf{B}$ and thus, $dE_k\mathbf{V}^* = d\mathbf{V}^*$. If $\mathbf{R}$ is $(A, E, \mathbf{B})$-inv., then $\mathbf{R} \subset \mathbf{V}^*$, and therefore $\mathbf{R} \cap \mathbf{Ker}\, \mathbf{E}_k = \mathbf{0}$. □

Lemma 5.3 will be instrumental in showing the following fact.

THEOREM 5.3. *A subspace* $\boldsymbol{R}$ *with* $d\mathbf{R} = q \geqq 1$ *is a generalized reachability subspace if and only if, given any symmetric set* Λ *of* q *complex numbers, there exist linear maps* F *and* K *such that we have the following*:

(1) (E_k, A_F) *is regular*;

(2) $A_F\mathbf{R} \subset E_k\mathbf{R}$;

(3) $\sigma(E_k|\mathbf{R}, A_F|\mathbf{R}) = \Lambda$.

Proof. Let $\mathbf{R}$ be a generalized reachability subspace. Choose K as in Lemma 5.3. Then $\mathbf{R}$ is a reachability subspace of (E_k, A, B) satisfying $\mathbf{R}\ \mathbf{Ker}\, \mathbf{E}_k = \mathbf{0}$. The result now follows from Corollary 5.1.

Now, suppose that there exist F and K to satisfy (1)-(3). As $q = \text{card}\ \sigma(E_k|\mathbf{R}, A_F|\mathbf{R}) \leqq \text{rank}\,(E_k|\mathbf{R}) \leqq q$, we have $\text{rank}\,(E_k|\mathbf{R}) = q$, that is, $\text{Ker}\,(E_k/\mathbf{R}) = 0$. Thus, $\mathbf{Ker}\, \mathbf{E}_k \cap \boldsymbol{R} = 0$. Then, by Lemma 5.1, $\mathbf{R}$ is a reachability subspace of (E_k, A, B). But then $\mathbf{R}$ is clearly a reachability subspace of (E, A, B) also. □

Indeed, it is also possible to prove a much stronger version of Theorem 5.3. However, as the proof that we have is too long to be reproduced here, we would rather state it as a conjecture.

CONJECTURE. *A subspace* $\mathbf{R}$ *with* $d\mathbf{R} = \geqq 1$ *is a generalized reachability subspace if and only if, given any integer* r, $0 \leqq r \leqq q$, *and any symmetric set* Λ *of* r *complex numbers, there exist linear maps* F *and* K *such that*:

(1) (E_k, A_F) *is regular*;

(2) $A_F\mathbf{R} \subset E_k\mathbf{R}$;

(3) $\sigma(E_k|\mathbf{R}, A_F|\mathbf{R}) = \Lambda$.

The conjecture above states that some or all of the eigenvalues of the closed-loop system restricted to R can be shifted to infinity if so desired.

6. Extension to nonregular systems. We emphasize that, although regularity of the open-loop system (1.1) has been a standing assumption throughout the paper, this choice has not been motivated by the dependence of the results on the regularity condition. Rather, we have been trying to simplify the exposition that has already been complicated enough by the condition of the closed-loop regularity. Now, assume that (1.1) is not regular, but is regularizable as defined in [20]. That is to say, suppose that $(E, A + BF_0, B)$ is a regular system for some linear map F_0. Suppose $\mathbf{R}$ is an $(A, E, \mathbf{B})$-inv. subspace that also satisfies GARSA given by (4.1). Then, $(A + BF_0)\mathbf{R} \subset E\mathbf{R} + \mathbf{B}$ and $\mathbf{R}$ also satisfies (4.1) when the recursion is performed with $(A + BF_0)$ rather than A. Thus, $\mathbf{R}$ is a reachability subspace of $(E, A + BF_0, B)$, and there exist linear maps F_1 and G so that $\mathbf{R}$ becomes the reachable subspace $\langle E, A + BF_0 + BF_1 | BG\rangle$ of a regular closed-loop system $(E, A + BF_0 + BF_1, BG)$. Letting $F = F_0 + F_1$, we realize that even when (1.1) is not regular, if $\mathbf{R}$ is an $(A, E, \mathbf{B})$-inv. subspace satisfying GARSA, then there exist linear maps F and G so that $R = \langle E, A + BF | \mathbf{BG}\rangle$.

7. Conclusions. It has been shown that the concept of reachability subspaces can be generalized to encompass singular systems also. This was accomplished by demonstrating that if (and only if) a given subspace $\mathbf{R}$ is an $(A, E, \mathbf{B})$-inv. (generalized)

almost-reachability subspace, then there exist two linear maps F and G such that $\mathbf{R}$ is the reachable subspace of a regular closed-loop system $(E, A+BF, BG)$. Instrumental in handling the complexities caused by the condition of closed-loop regularity was the notion of "regular friends" of an $(A, E, \mathbf{B})$-inv. subspace. We emphasize that our approach uses only proportional feedback, as opposed to the constant-ratio proportional-plus-derivative feedback used in [23].

A subspace $\mathbf{C} \subset E\mathbf{X}$ was defined to be a generalized controllability subspace if $\mathbf{C} = E\mathbf{R}$ for some generalized reachability subspace (in which case C becomes the controllable subspace of a regular system $(E, A+AB, BG)$ for some F and G). Spectral assignability properties of generalized controllability subspaces (as well as those of generalized reachability subspaces) were also discussed.

To simplify the presentation, we assumed a regular (i.e., $\det(sE-A) \not\equiv 0$) system. However, all the results were finally extended to the nonregular case.

Appendix.

Proof of Theorem 2.3. Let $K : \mathbf{X} \to \mathbf{U}$ be as constructed in the proof of Lemma 5.3. Then K satisfies $\mathbf{Ker}\, \mathbf{E}_k \cap \mathbf{V}^* = \mathbf{0}$ and $E\mathbf{V}^* + \mathbf{B} = E_k\mathbf{V}^*$. Now, it is easy to mimic the proof of Theorem 5.6 in [27] to show that $\mathbf{S}_*(\mathbf{V}^*) := \mathbf{S}_*(E, A, \mathbf{B}; \mathbf{V}^*)$ is $(A, E, \mathbf{B})$-inv. Then an $F: \mathbf{X} \to \mathbf{U}$ satisfying $A_F\mathbf{S}_*(\mathbf{V}^*) \subset E_k\mathbf{S}_*(\mathbf{V}^*)$ can be found. Note that $A_F\mathbf{V}^* \subset A\mathbf{V}^* + \mathbf{B} \subset E\mathbf{V}^* + \mathbf{B} = E_k\mathbf{V}^*$. This $\mathbf{V}^*$ is $(A_F, E_k, \mathbf{0})$-inv. and is, therefore, included in the initial manifold of (E_k, A_F). On the other hand, as $(A_F, E_k, \mathbf{0})$ invariance clearly implies $(A, E, \mathbf{B})$ invariance, the initial manifold of (E_k, A_F) is also included in $\mathbf{V}^*$. That is to say, $\mathbf{V}^*$ is the initial manifold of (E_k, A_F). Note that Theorem 2.1 and the fact that $\mathbf{Ker}\, \mathbf{E}_k \cap \mathbf{V}^* = \mathbf{0}$ imply that (E_k, A_F) is regular. Let $\mathbf{S}_k^0$ denote the final manifold $S_*(E_k, A_F, \mathbf{0}; \mathbf{X})$ of (E_k, A_F).

If S_k is defined by $\mathbf{S}_{k+1} = E^{-1}(A\mathbf{S}_k + \mathbf{B})$; $\mathbf{S}_0 = \mathbf{0}$, then $\mathbf{S}_0\, \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0$. If $\mathbf{S}_k\mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0$ is assumed, then it follows that

$$
\begin{aligned}
\mathbf{S}_{k+1} &\subset E^{-1}\{A\mathbf{S}_*(\mathbf{V}^*) + A\mathbf{S}_*^0 + \mathbf{B}\} \\
&\subset E_k^{-1}\{A_F\mathbf{S}_*(\mathbf{V}^*) + A_F\mathbf{S}_*^0 + \mathbf{B}\} \\
&\subset E_k^{-1}\{E_k\mathbf{S}_*(\mathbf{V}^*) + E_k\mathbf{S}_*^0 + \mathbf{B}\} \\
&\subset E_k^{-1}\{E_k\mathbf{S}_*(\mathbf{V}^*) + E_k\mathbf{S}_*^0 + \mathbf{B} \cap \mathbf{E}_k\mathbf{V}^*\} \quad \text{(because } B \subset E_k\mathbf{V}^*\text{)} \\
&\subset \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0 + \mathbf{V}^* \cap E_k^{-1}\mathbf{B} + \mathbf{Ker}\, \mathbf{E}_k \quad \text{(by Properties 2.1 and 2.2)} \\
&\subset \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_k^0 \quad \text{(because } \mathbf{Ker}\, \mathbf{E}_k \subset \mathbf{S}_*^0 \text{ and } \mathbf{V}^* \cap E_k^{-1}\mathbf{B} \subset \mathbf{S}_*(\mathbf{V}^*)\text{)}.
\end{aligned}
$$

Thus, $\mathbf{S}_k \subset \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0$ for all k, and therefore, $\lim_k \mathbf{S}_k := \mathbf{S}_* \subset \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0$. Then

$$
\begin{aligned}
\mathbf{S}_* \cap \mathbf{V}^* &\subset (\mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0) \cap \mathbf{V}^* \\
&\subset \mathbf{S}_*(\mathbf{V}^*) + \mathbf{S}_*^0 \cap \mathbf{V}^* \quad \text{(by Property 2.1)}.
\end{aligned}
$$

As (E_k, A_F) is regular and $\mathbf{V}^*$ and $\mathbf{S}_*^0$ are the initial and final manifolds of (E_k, A_F), respectively, $\mathbf{S}_*^0 \cap \mathbf{V}^* = \mathbf{0}$ by Theorem 2.1. Consequently, $\mathbf{S}_* \cap \mathbf{V}^* \subset \mathbf{S}_*(\mathbf{V}^*)$. It is also trivially true that $\mathbf{S}_*(\mathbf{V}^*) \subset \mathbf{S}_* \cap \mathbf{V}^*$. Then $\mathbf{S}_*(\mathbf{V}^*) = \mathbf{S}_* \cap \mathbf{V}^*$, and the proof is complete. □

REFERENCES

[1] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, San Francisco, 1980.

[2] J. D. COBB, *Descriptor variable and generalized singularly perturbed systems: a geometric approach*, Ph.D. thesis, Department of Electrical Engineering, University of Illinois, Chicago, IL, 1980.

[3] J. D. COBB, *Controllability, observability and duality in singular systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 1076–1082.
[4] J. D. COBB, *Feedback and pole placement in descriptor variable systems*, Internat. J. Control, 33 (1981) pp. 1135–1146.
[5] F. R. GANTMACHER, *Theory of Matrices*, Chelsea, New York, 1960.
[6] L. KRONECKER, *Algebraische Reduction der schaaren bilinearer Forman*, S.-B. Akad. Berlin, 1890, pp. 763–776.
[7] V. KUČERA AND P. ZAGALAK, *Fundamental theorem of state feedback for singular systems*, Automatica, 24 (1988), pp. 653–658.
[8] F. L. LEWIS AND K. ÖZCALDIRAN, *Reachability and controllability for descriptor systems*, in Proc. 27th Midwest Symposium on Circuits and Systems, Morgantown, WV, 1984, pp. 676–678.
[9] F. L. LEWIS, *A survey of singular systems*, Circuits Systems Signal Process. (special issue on Semistate Systems), 5 (1986), pp. 3–36.
[10] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312–321.
[11] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473–480.
[12] M. Malabre, *More geometry about singular systems*, Proc. IEEE Conference on Decision and Control, Los Angeles, CA, December 1987, pp. 1138–1139.
[13] R. MUKUNDAN AND W. DAYAWANSA, *Feedback control of singular systems—proportional and derivative feedback of the state*, Internat. J. Systems Sci., 14 (1983), pp. 615–632.
[14] R. W. NEWCOMB, *The semistate description of nonlinear time-variable circuits*, IEEE Trans. Circuits Systems, 28 (1981), pp. 62–71.
[15] ———, *Semistate design theory: binary and swept hysteresis*, Circuits Systems Signal Process. 1 (1982), pp. 203–216.
[16] R. W. NEWCOMB AND N. EL-LEITHY, *Semistate description of an* MOS *neural-type cell*, Proc. Mathematical Theory of Networks and Systems Conference, Phoenix, AZ, June 1987.
[17] K. ÖZCALDIRAN, *Control of descriptor systems*, Ph.D. thesis, School of Electrical Engineering, Georgia Institute of Technology, Atlanta, GA, 1985.
[18] K. ÖZCALDIRAN AND F. L. LEWIS, *A geometric approach to eigenstructure assignment for singular systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 629–632.
[19] K. ÖZCALDIRAN, *A geometric characterization of the reachable and the controllable subspaces of descriptor systems*, Circuits Systems Signal Process. (special issue on Semistate Systems), 5 (1986), pp. 37–48.
[20] K. ÖZCALDIRAN AND F. L. LEWIS, *On the regularizability of singular systems*, IEEE Trans. Automat. Control, to appear.
[21] H. H. ROSENBROCK AND A. C. PUGH, *Contributions to a hierarchical theory of systems*, Internat. J. Control, 19 (1974), pp. 845–867.
[22] J. M. SCHUMACHER, *Algebraic characterization of almost invariance*, Internat. J. Control, 38 (1983), pp. 107–124.
[23] M. A. SHAYMAN AND Z. ZHOU, *Feedback control and classification of generalized linear systems*, IEEE Trans. Automat. Control, (1987), pp. 483–494.
[24] G. C. VERGHESE, *Infinite-frequency behavior in generalized dynamical systems*, Ph.D. thesis, Department of Electrical Engineering, Stanford University, Stanford, CA, 1978.
[25] K. WEIERSTRASS, *Zur Theorie der bilinearen und quadratischen Former*, Monatsh. Akad. Wiss. Berlin, 1867, pp. 310–338.
[26] J. H. WILKINSON, *Linear differential equations and Kronecker's canonical form*, in Recent Advances in Numerical Analysis, C. de Boor and G. Golub, eds., Academic Press, New York, 1978, pp. 231–265.
[27] K. T. WONG, *The eigenvalue problem* $\lambda Tx + Sx$, J. Differential Equations, 16 (1974), pp. 270–280.
[28] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1979.
[29] J. C. WILLEMS, *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.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 511-526, May 1989

004

H_∞ INTERPOLATION IN SYSTEMS WITH COMMENSURATE INPUT LAGS*

GILEAD TADMOR†

Abstract. H_∞ optimal control and system design problems are known to mathematically reduce to the framework of operator interpolation. Here a solution scheme in the context of systems with multiple input lags is provided. The main result, a linear algebraic characterization of eigenvalue/eigenfunction pairs, can be utilized also in finite-dimensional approximations of the associated Hankel Operator. A particular feature of the analysis is the heavy reliance on time-domain techniques and utilization of explicit time-domain properties of the transfer function.

Key words. H_∞ interpolation, input delays, time domain analysis, eigenvalues, eigenfunctions

AMS(MOS) subject classifications. 47A20, 30D55, 49A34

1. Introduction. Our note is part of a recent effort to develop a theory of H_∞ optimization for distributed parameter systems. Strong motivations come from control and system engineering. H_∞ theory addresses a wide spectrum of worst-case analysis design and control issues, including robust control, weighted sensitivity minimization, model matching, tracking error attenuation, gain-phase margins, and the like (see [12], [13], [17], [22], [23] for overviews and references). Problems involving distributed parameter systems that are currently of great interest, such as the robust design of large space structures, or the control of flexible robot arms, naturally fall in its domain. The general difficulty is, of course, in handling the inherent infinite dimensionality of distributed parameter systems.

The central component in an H_∞ optimization procedure is the (approximate or exact) solution of an operator interpolation problem. (An account of the relation of operator interpolation to H_∞ optimization can be found in the manuscript [12]. The technical details of the problem are given in § 2.) In the context of distributed parameter systems, the operator to be interpolated, say T, becomes infinite-dimensional, which renders the problem quite difficult. First attempts to solve it were restricted, therefore, to case studies [3]–[6] of SISO (scalar) systems with single pure delay; these are perhaps the simplest distributed parameter systems. Nonetheless, the analysis revealed considerable complexity already in this case.

Delay systems are of interest for their own virtue. But they also serve as relatively simple realizations of input-output maps, in some systems that are actually governed by partial differential equations (PDEs), and distributed control mechanisms. (Examples are linear elements subject to tension, compression and torsion, with boundary control and observations. These systems are internally modeled by the wave equation; yet, their input-output transfer function is that of a delay system.) To my knowledge, the present investigation is the first beyond the initial studies mentioned above to provide a solution scheme in the context of an infinite-dimensional system. It treats the entire class of systems with multiple (commensurate) input lags, which seems a natural follow-up. Our technique readily generalizes to also cover certain systems with distributed input delays [21]. Hopefully, it will provide insight into H_∞ optimization problems in other interesting classes.

* Received by the editors May 12, 1986; accepted for publication (in revised form) May 24, 1988. This paper is a revised version of an MIT Technical Report, LIDS-P-1546.

† Program in Mathematical Sciences, The University of Texas at Dallas, Richardson, Texas 75083-0688.

A special feature of the developments in this paper is that they are based on a predominately time-domain analysis. We derive and work with time-domain formulae for the operator T and for its adjoint T^*. We do so utilizing the time-domain meaning of the system's transfer function and some elementary semigroup theory. In particular, we do not use an inner-outer factorization of the transfer function, as is the common practice.

Our main result, Observation 4.1, is a workable (linear algebraic) characterization of eigenvalue/eigenfunction pairs for T^*T. These pairs are known to play a key role in suboptimal solutions, via finite-dimensional approximations of T (see, e.g., [18]), and in optimal solutions, as we describe in § 2, below.

Indeed, studies that became available following the preparation of the first version of this paper [7]–[11], [24] concentrate also on spectral analysis of the operator T, with some beautiful results. In particular, [11] provides an algebraic characterization of eigenvalue/eigenfunction pairs in a very general setting. In point of fact, we were motivated by the discussion in [24] to improve certain developments from the original version of this paper. We shall make note of these improvements during the discussion.

The paper is organized as follows: the problem statement and some preliminary results are given in § 2. Time domain formulae for T and for T^* are derived in § 3. In § 4 we characterize eigenvalue/eigenfunction pairs for T^*T, and conclude, in § 5, with an illustrating example.

2. Preiiminaries. H_2 and H_∞ are the usual Hardy spaces of analytic functions on the open right half complex plane, with L_2 and L_∞ boundary values (respectively) on the imaginary axis (the frequency domain). Since the Laplace transform defines an isometry between $L_2[0,\infty)$ and H_2, we shall not distinguish between a function in $L_2[0,\infty)$ and its transform in H_2 by some particular notation (e.g., by "$\wedge$" or "$\vee$"). Thus if $f\in H_2$ then $f(t)$ will be the time-domain function, $f(s)$ will be its Laplace transform, defined over the right half-plane, and $f(j\omega)$ will be the boundary function of the latter, over the frequency domain. A good source on Hardy spaces, inner and outer functions, etc., is [14].

A star will denote the adjoint of an operator (e.g., T^*, in particular, if A is a matrix then A^* is the transposed complex conjugate of A), and the conjugate $f^*(s)=\overline{f(-\bar{s})}$ of an analytic function.

The following is a general setup of the operator interpolation problem. Given are a rational function $w(s)\in H_\infty$, and an inner function $m(s)$. Let K be the orthogonal complement of $m(s)H_2$ in H_2 ($K=H_2\ominus mH_2$), let π be the orthogonal projection of H_2 onto K, and let $T: K\to K$ be the bounded linear operator formed by compression of multiplication by $w(s)$ to the space K. Namely, $Tf=\pi(wf)$ for $f\in K$. A function $w_0\in H_\infty$ *interpolates* T if it satisfies (i) $Tf=\pi(w_0f)$ and (ii) $\|w_0\|_\infty=\|T\|$ (equals the induced operator norm of T). We look for an interpolating function.

Notice that, by definition of T, the function $w(s)$ satisfies (i), but that in general it can only be expected that $\|w\|_\infty\geqq\|T\|$. Condition (ii) is therefore an optimality requirement on the family of functions $w_0\in H_\infty$ that define T via (i).

Details on the way to reduce an H_∞ system optimization problem to the framework of an interpolation problem can be found in [12]. Let us just mention here that $m(s)$ is the inner part of the transfer function $G(s)$ of a system at hand, and that $w(s)$ reflects a design objective; e.g., a weight function in the context of weighted sensitivity minimization. Consequently, optimal compensators are computed in terms of interpolating functions.

Let us also mention that substituting $G(s)$ for $m(s)$ in the definition of the subspace K will have no effect on that definition. The reason is that for an outer function $\psi(s) \in H_2$, the linear manifold $\psi(s)H_2$ is dense in H_2.

Here we are interested in systems with multiple input delays. We assume, therefore, that $G(s) = P(e^{-s})$, where $P(z) = \sum_{i=0}^{n} a_i z^i$ is a polynomial. For technical simplicity we also assume the following: $w(s)$ has only first-order poles; using the notation $w(s) = \eta + \sum_{i=1}^{m} \alpha_i/(s+\beta_i)$; $w(s) = 0$ implies Re $s < 0$ (i.e., $w(s)$ is of minimum phase); and $P(z) = 0$ implies $|z| < 1$.

Sarason [20] has established the existence of an interpolating function. The following observation of Sarason is generally used as a starting point for computing it.

LEMMA 2.1 [20, Prop. 5.1]. *Suppose the operator T has a maximal function $f \in K$. Then the unique interpolating function is $w_0(s) = Tf(s)/f(s)$. The function $w_0(s)$ is all-pass: $|w_0(jw)| = \|T\|$ almost everywhere.*

It follows from Weyl's lemma [16, pp. 32, 295] that $\|T^*T\|$ is equal to the spectral radius $\rho(T^*T)$. (We shall justify the use of Weyl's lemma in the proof of Observation 2.2 in § 4.) We are therefore led to compute $\rho(T^*T)$ and check whether it is an eigenvalue of T^*T. For if it is, then the associated eigenfunction is the desired maximal function for T. Via Lemma 2.1, that function provides the solution to the interpolation problem.

In § 4 we find a parameterized family of $2m \times 2m$ matrices $\Omega(\lambda^2)$, $\lambda^2 \geqq 0$, such that λ^2 is an eigenvalue if and only if $\det \Omega(\lambda^2) = 0$. Associated eigenfunctions are then easily recovered from zero vectors of $\Omega(\lambda^2)$. This characterization is to be used in searching for maximal eigenvalue/eigenfunction pairs. The following observation restricts the domain of the numerical search. (The proof is deferred to § 4.)

OBSERVATION 2.2. (1) *If for some $\omega_0 > 0$ there holds $|w(j\omega)| > |\eta|$ almost everywhere for $|\omega| > \omega_0$, then a maximal eigenvalue exists.*

(2) *If a maximal eigenvalue exists, it is situated in the interval $[|\eta|^2, \|w\|_\infty^2]$.*

(3) *If $|w(j\omega)| \leqq \eta$ for all ω, then unless $w(s) \equiv \eta$ (then $T = \eta I$), a maximal eigenvalue does not exist. Yet, then $w(s)$ itself is an interpolating function.*

3. The time-domain setup. Our first task is to provide a tangible description of the subspace $K = H_2 \ominus P(e^{-s})H_2$, and a workable formula for the projection π. We do so in the time domain, identifying H_2 with $L_2[0, \infty)$.

OBSERVATION 3.1. *The subspace K consists of all the solutions to the difference equation*

$$\sum_{i=0}^{\infty} \bar{a}_i f(t+i) = 0 \quad \text{for } t \in [0, \infty), \tag{3.1}$$

that satisfy $f|_{[0,n]} \in L_2[0, n]$.

Proof. Let $\mathbf{P}: H_2 \to H_2$ be the multiplication operator associated with the function $P(e^{-s})$. By Fredholm's alternative, $K = \ker \mathbf{P}^* \subset L_2[0, \infty)$. Provided that $f \in L_2[0, \infty)$, equation (3.1) explicitly rewrites the equality $\mathbf{P}^* f = 0$. Hence, we have the necessity of the conditions on f.

To establish sufficiency it has to be shown that a solution of (3.1), with an initial trajectory in $L_2[0, n]$, must belong to $L_2[0, \infty)$. Indeed, let $S(t)$ be the c_0-semigroup on $L_2[0, n]$, which shifts along solutions. That is, if f is a solution with $f|_{[0,n]} = \xi \in L_2[0, n]$, then $[S(t)\xi](\tau) = f(t+\tau)$ for $\tau \in [0, n]$. The characteristic equation of the infinitesimal generator of $S(\cdot)$ is $G^*(s) = P(e^{\bar{s}}) = 0$. Our assumption ($P(z) = 0 \Rightarrow |z| < 1$, in § 2) is that solutions of this equation all lie strictly within the open left half-plane.

Thus it follows, from the standard theory of delay and difference equations (see, e.g., [15]), that $\|S(t)\| \leqq Me^{-\varepsilon t}$ for some positive constants M and ε. Hence the claim is true.

For future use we add the following notation: $E = S(n)$ and $Q = \sum_{i=0}^{\infty} E^{*i}E^i$ are operators on $L_2[0, n]$. (The stability of $S(\cdot)$ assures that Q is well defined.) Given a function f on $[0, \infty)$, and $t \geqq 0$, the symbol f^t stands for the function on $[0, n]$, defined by $\tau \to f(t+\tau)$.

COROLLARY 3.2. *The orthogonal projection $\pi : L_2[0, \infty) \to K$ is given by the formula*

$$[\pi f]^{ln} = E^l Q^{-1} \sum_{i=0}^{\infty} E^{*i} f^{in}, \qquad l = 0, 1, 2, \cdots. \tag{3.2}$$

Proof. Taking the right-hand side of (3.2) as a definition of π, we shall now show that it is the orthogonal projection onto K. That is, that (i) Im $\pi \subset K$, (ii) $\pi|_K = I|_K$, and (iii) $\pi^* = \pi$.

(i) As in the previous proof, $\|E^{*i}\| \leqq Me^{-\varepsilon in}$. Hence, the series on the right-hand side of (3.2) converges for $f \in L_2[0, \infty)$, and $[\pi f]^0 \in L_2[0, n]$. The definition of E implies that πf satisfies (3.1), and by Observation 3.1 it is a function in K.

(ii) If $f \in K$ then $f^{in} = E^i f^0$, whereby

$$\begin{aligned}[\pi f]^{ln} &= E^l Q^{-1} \sum_{i=0}^{\infty} E^{*i} E^i f^0 \\ &= E^l Q^{-1} Q f^0 = E^l f^0 = f^{ln}.\end{aligned}$$

So $\pi|_K = I|_K$.

(iii)

$$\begin{aligned}\langle \pi f, g\rangle_{L_2[0,\infty)} &= \sum_{l=0}^{\infty} \langle \pi f^{ln}, g^{ln}\rangle_{L_2[0,n]} \\ &\sum_{l=0}^{\infty} \left\langle E^l Q^{-1} \sum_{i=0}^{\infty} E^{*i} f^{in}, g^{ln} \right\rangle_{L_2[0,n]} \\ &= \sum_{i=0}^{\infty} \left\langle f^{in}, E^{in} Q^{-1} \sum_{l=0}^{\infty} E^{*ln} g^{ln} \right\rangle_{L_2[0,n]} \\ &= \langle f, \pi g\rangle_{L_2[0,\infty]}.\end{aligned}$$

COROLLARY 3.3. *Let X be the space $L_2[0, n]$ when endowed with the inner product $\langle \cdot, \cdot \rangle_X = \langle \cdot, Q\cdot \rangle_{L_2[0,n]}$. Then K and X are isometric, and the isometry is defined by the mapping $\iota : X \to K$, where*

$$[\iota\xi]^{ln} := E^l \xi \quad \text{for } l = 0, 1, 2, \cdots.$$

Proof. Following from Observation 3.1, we have that the mapping ι defines an isomorphism between X and K, and $\iota^{-1} f = f^0$ for $f \in K$. The observation that ι is isometric follows from

$$\begin{aligned}\|\iota\xi\|_K^2 &= \langle \iota\xi, \iota\xi\rangle_{L_2[0,\infty)} \\ &= \sum_{i=0}^{\infty} \langle [\iota\xi]^{in}, [\iota\xi]^{in}\rangle_{L_2[0,n]} \\ &= \sum_{i=0}^{\infty} \langle E^i\xi, E^i\xi\rangle_{L_2[0,n]} \\ &= \langle \xi, Q\xi\rangle_{L_2[0,n]} = \|\xi\|_X^2.\end{aligned}$$

So far we have essentially used only the fact that the difference equation $\mathbf{P}^* f = 0$ gives rise to an exponentially stable semigroup. Taking into account the commensurate structure of the delay operator $P(e^{-s})$, we discover that the operators E, E^*, and Q (indeed, the key elements in our formulae) have very simple, finite-dimensional structures, as follows: Set

$$\tilde{E} = -\begin{bmatrix} \bar{a}_n & 0 & \cdots & & 0 \\ \bar{a}_{n-1} & \bar{a}_n & 0 & \cdots & 0 \\ \vdots & & & & \\ \bar{a}_1 & \cdots & & \bar{a}_{n-1} & \bar{a}_n \end{bmatrix}^{-1} \begin{bmatrix} \bar{a}_0 & \bar{a}_1 & \cdots & & \bar{a}_{n-1} \\ 0 & \bar{a}_0 & \bar{a}_1 & \cdots & \bar{a}_{n-2} \\ \vdots & & & & \\ 0 & \cdots & & 0 & \bar{a}_0 \end{bmatrix},$$

and, given a (scalar) function $\xi \in X$, define an n-vector function $\tilde{\xi}$, on $[0, 1)$, by

$$\tilde{\xi}(\tau) = \begin{bmatrix} \xi(\tau) \\ \xi(\tau+1) \\ \vdots \\ \xi(\tau+n-1) \end{bmatrix} \quad \text{for } \tau \in [0, 1).$$

OBSERVATION 3.4. *$\zeta = E\xi \Leftrightarrow \tilde{\zeta} = \tilde{E}\tilde{\xi}$ and $\zeta = E^*\xi \Leftrightarrow \tilde{\zeta} = \tilde{E}^*\tilde{\xi}$.*

Proof. The observation follows directly from (3.1).

Consequently, $\tilde{Q} = \sum_{i=0}^{\infty} \tilde{E}^{*i}\tilde{E}^i$ is a matrix representation of the operator Q. It can be computed in a finite process, as a solution of the Lyapunov equation $\tilde{Q} - \tilde{E}^*\tilde{Q}\tilde{E} = \tilde{I}$ (see [1] for detail). Henceforth we shall use these matrix representations, and in particular, the associated matrix version $\tilde{\pi}$ of the projection onto K.

We turn to the operators T and T^*. In view of Corollary 3.3, we shall interpret them as defined over the space X. The appropriate definition of T is thereby $T\xi = [\pi(w * \iota\xi)]^0$, where $*$ stands for the convolution in $L_2[0, \infty)$ and $w(t)$ is given by

$$w(t) = \eta\delta(t) + \sum_{i=1}^{m} \alpha_i e^{-\beta_i t}, \qquad t \geqq 0$$

(with $\operatorname{Re} \beta i > 0$, since $w(s) \in H_\infty$).

Since for $\xi \in X$ we have $\delta * \eta\xi = \eta\xi$ and $[\pi \circ \iota\eta\xi]^0 = \eta\xi$, it remains to compute concrete formulae for terms of the form $[\pi(e^{-\beta \cdot} * \iota\xi)]^0$, where $\operatorname{Re} \beta > 0$. Invoking Observation 3.4, a straightforward computation yields

$$\widetilde{[e^{-\beta \cdot} * \iota\xi]^{in}}(\tau) = \tilde{E}^i \int_0^\tau e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\, d\theta$$

$$+ \left(\tilde{F}_0\tilde{E}^i + \tilde{F}_1 \sum_{k=0}^{i-1} e^{(k-i)n\beta}\tilde{E}^k \right) \int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\, d\theta, \qquad \tau \in [0, 1) \tag{3.3}$$

where $\sum_{k=0}^{i-1} := 0$ for $i = 0$, and the matrices $\tilde{F}_0 = \tilde{F}_0(\beta) = [(f_0)_{pq}]$ and $\tilde{F}_1 = \tilde{F}_1(\beta) = [(f_1)_{pq}]$ are as follows:

$$(f_0)_{pq} = \begin{cases} e^{(q-p)\beta} & \text{for } n \geqq p > q \geqq 1, \\ 0 & \text{for } 1 \leqq p \leqq q \leqq n, \end{cases}$$

and

$$(f_1)_{pq} = e^{(q-p)\beta} \quad \text{for } p, q = 1, 2, \cdots, n.$$

For simplicity we assume that $e^{-n\beta}$ is not an eigenvalue of $\tilde{E}$ and rewrite (3.3) as

$$
\begin{aligned}
(3.3^0)\qquad \cdots &= \tilde{E}^i \int_0^\tau e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta \\
&\quad + ((\tilde{F}_0 - \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1})\tilde{E}^i + \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1}\,e^{-in\beta}) \int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta.
\end{aligned}
$$

Now

$$
\begin{aligned}
[\pi(\widetilde{e^{-\beta\cdot}} * \iota\xi)]^0(\tau) &= \tilde{Q}^{-1}\sum_{i=0}^{\infty} \tilde{E}^{*i}[\widetilde{e^{-\beta\cdot}} * \iota\xi]^{in}(\tau) \\
&= \tilde{Q}^{-1}\sum_{i=0}^{\infty} \tilde{E}^{*i}\tilde{E}^i \int_0^\tau e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta \\
&\quad + \tilde{Q}^{-1}\sum_{i=0}^{\infty} \tilde{E}^{*i}((\tilde{F}_0 - \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1})\tilde{E}^i \\
&\quad + \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1}\,e^{-in\beta}) \\
&\quad \cdot \int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta \\
&= \tilde{Q}^{-1}\tilde{Q}\int_0^\tau e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta \\
&\quad + \tilde{Q}^{-1}\left(\left(\sum_{i=0}^{\infty} \tilde{E}^{*i}(\tilde{F}_0 - \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1})\tilde{E}^i\right)\right. \\
&\quad \left. + (\tilde{I}-\tilde{E}^*\,e^{-n\beta})^{-1}\tilde{F}_1(I-\tilde{E}\,e^{n\beta})^{-1}\right) \cdot \int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta \\
&= \int_0^\tau e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta + \tilde{G}\int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta
\end{aligned}
\tag{3.4}
$$

where

$$
\tilde{G} = \tilde{G}(\beta) = \tilde{Q}^{-1}(\tilde{Y} + (\tilde{I}-\tilde{E}^*\,e^{-n\beta})^{-1}\tilde{F}_1(I-\tilde{E}\,e^{n\beta})^{-1})
$$

and

$$
\tilde{Y} = \sum_{i=0}^{\infty} \tilde{E}^{*i}(\tilde{F}_0 - \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1})\tilde{E}^i.
$$

Note that by assumption ($\operatorname{Re}\beta > 0$) the matrix $\tilde{I}-\tilde{E}^*\,e^{-n\beta}$ is indeed invertible, and that $\tilde{Y}$ can be computed by solving the Lyapunov equation $\tilde{Y}-\tilde{E}^*\tilde{Y}\tilde{E} = \tilde{F}_0 - \tilde{F}_1(\tilde{I}-\tilde{E}\,e^{n\beta})^{-1}$. Substituting β_i for β and $\tilde{G}_i = \tilde{G}(\beta_i)$ for G, we thus have observation 3.5.

OBSERVATION 3.5. *Interpreted as defined on X, the operators T and T^* are given by the formulae*

$$
\begin{aligned}
[\widetilde{T\xi}](\tau) &= \eta\tilde{\xi}(\tau) + \sum_{i=1}^{m} \alpha_i \int_0^\tau e^{\beta_i(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta + \sum_{i=1}^{m} \alpha_i G_i \int_0^1 e^{\beta_i(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta, \\
[\widetilde{T^*\zeta}](\tau) &= \bar{\eta}\tilde{\zeta}(\tau) + \sum_{i=1}^{m} \bar{\alpha}_i \int_\tau^1 e^{\bar{\beta}_i(\tau-\theta)}\tilde{\zeta}(\theta)\,d\theta + \sum_{i=1}^{m} \bar{\alpha}_i\tilde{Q}^{-1}\tilde{G}_i^*\tilde{Q}\int_0^1 e^{\bar{\beta}_i(\tau-\theta)}\tilde{\zeta}(\theta)\,d\theta \\
&= \bar{\eta}\tilde{\zeta}(\tau) - \sum_{i=1}^{n} \bar{\alpha}_i \int_0^\tau e^{\bar{\beta}_i(\tau-\theta)}\tilde{\zeta}(\theta)\,d\theta + \sum_{i=1}^{n} \bar{\alpha}_i[\tilde{I}+\tilde{Q}^{-1}\tilde{G}_i^*\tilde{Q}]\int_0^1 e^{\bar{\beta}_i(\tau-\theta)}\tilde{\zeta}(\theta)\,d\theta
\end{aligned}
$$

for $\tau\in[0,1]$, and where $\tilde{G}_i^*$ is the usual matrix adjoint.

The proof is obvious once we notice that the adjoint of the operator $G: X \to X$ has the matrix representation $\tilde{Q}^{-1}\tilde{G}^*\tilde{Q}$.

In what follows, it will be useful to work also with scalar formulae for T and T^*, utilizing this next observation on the exact structure of the matrices $\tilde{G}_i$ and $\tilde{I} + \tilde{Q}^{-1}\tilde{G}_i^*\tilde{Q}$.

OBSERVATION 3.6. *Let $\tilde{G}$ be as in formula* (3.4). *Then three exist n-vectors $\tilde{a}$, $\tilde{b}$, $\tilde{c}$, and $\tilde{d}$ such that*

$$\tilde{G}(\beta) - \tilde{F}_0(\beta) = ab^* \quad \text{and} \quad \tilde{I} + \tilde{Q}^{-1}\tilde{G}^*(\beta)\tilde{Q} + \tilde{F}_0(-\bar{\beta}) = \tilde{c}\tilde{d}^*.$$

Proof. For the moment we return to a frequency domain discussion, motivated by [24, § 1.2]. Let π_+ and π_- be the orthogonal projections from $L_2(j\mathfrak{R})$ onto H_2 and $H_2^\perp$, respectively, and recall that $m(s)$ is the inner part of $G(s) = P(e^{-s})$. It is very easy to see that the frequency domain formula for the projection $\pi: H_2 \to K$ is $\pi f(s) = m(s)\pi_-(m^*(s)f(s))$. Thus

$$\begin{aligned} \pi\left(\frac{1}{s+\beta}f(s)\right) &= m(s)\pi_-\left(m^*(s)\frac{1}{s+\beta}f(s)\right) \\ &= \frac{1}{s+\beta}f(s) - m(s)\pi_+\left(\frac{1}{s+\beta}m^*(s)f(s)\right) \end{aligned} \tag{3.5}$$

for $f \in K$.

Consider the second term on the right-hand side of (3.5). For $f \in K$, the function $m^*(s)f(s)$ belongs to $H_2^\perp$. It follows from the standard theory of Hankel operators (see, e.g., [12, Chap. 6]) that the operator that takes $h(s) \in H_2^\perp$ to $\pi_+(1/(s+\beta)h(s)) \in H_2$ is of rank one. We rewrite the conclusion in a vector form, and in the time domain

$$\pi(\widetilde{e^{-\beta\cdot}} * f) - \widetilde{e^{-\beta\cdot}} * f = \text{a first-order operator applied to } f \quad \text{for } f \in K. \tag{3.6}$$

Now we are almost done! Invoking formulae (3.3^0) (with $i = 0$) and (3.4), the left-hand side of (3.6) is equal to

$$\cdots = (\tilde{G} - \tilde{F}_0)\int_0^1 e^{\beta(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta.$$

So $\tilde{G} - \tilde{F}_0$ must be a matrix of rank one, namely, of the form $\tilde{a}\tilde{b}^*$.

The conclusion regarding the matrix $\tilde{I} + \tilde{Q}^{-1}\tilde{G}^*\tilde{Q}$ follows from a completely similar line of arguments, applied to the operator T^*.

COROLLARY 3.7. *Let $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{c}_i$, and $\tilde{d}_i$ correspond to the matrix $\tilde{G}_i$, $i = 1, \cdots, m$, as in the previous observation. Then*

$$\begin{aligned} T\xi(\tau) &= \eta\xi(\tau) + \sum_{i=1}^m \alpha_i \int_0^\tau e^{\beta_i(\theta-\tau)}\xi(\theta)\,d\theta \\ &\quad + \sum_{i=1}^m \alpha_i e^{-\beta_i\tau}\sum_{k=0}^{n-1} e^{\beta_i k}a_{ik}\chi_{[k,k+1)}(\tau)\int_0^1 e^{\beta_i\theta}\langle\tilde{b}, \tilde{\xi}(\theta)\rangle\,d\theta, \\ T^*\zeta(\tau) &= \bar{\eta}(\tau) - \sum_{i-1}^m \bar{\alpha}_i \int_0^\tau e^{\bar{\beta}_i(\tau-\theta)}\zeta(\theta)\,d\theta \\ &\quad + \sum_{i=1}^m \bar{\alpha}_i e^{\bar{\beta}_i\tau}\sum_{k=0}^{n-1} e^{-\bar{\beta}_i k}c_{ik}\chi_{[k,k+1)}(\tau)\int_0^1 e^{-\bar{\beta}_i\theta}\langle\tilde{d}, \tilde{\zeta}(\theta)\rangle\,d\theta \end{aligned}$$

for $\tau \in [0, n]$, and where $\chi_{[\varepsilon,\delta]}(\cdot)$ is the usual characteristic function of the interval $[\varepsilon, \delta]$. The scalars a_{ik} and c_{ik} are the $k+1$ components of the n-vectors $\tilde{a}$ and $\tilde{c}$, respectively.

Remark. Information on the order of zero of $P(z)$ at $z=0$ can help in simplifying some of the constructions above and facilitate the numerical computations suggested in the following section. The particular case $P(z)=z^n$ is the simplest: it stands for a single, pure input lag, which is the case studied in [2]–[6]. In those papers it is observed that then K is isometric to $L_2[0, n]$, and that T is the restriction of the Volterra operator of convolution with $w(t)$, to the interval $[0, n]$. Indeed, it turns out that if $a_0=a_1=\cdots=a_{n-1}=0$, and $a_n\neq 0$ (which is our case), then the matrix $\tilde{E}$ vanishes, $\tilde{Q}$ is simply the identity, and $\tilde{G}=\tilde{F}_0$, for all i.

In a more general case $P(z)$ will have a zero of order $1\leqq k\leqq n$ at the origin, i.e., $a_0=a_1=\cdots+a_{k-1}=0$, $a_k\neq 0$. This information can be utilized as follows. Denote $b_i=a_{i+k}$ for $i=0, 1, \cdots, n-k$, and define a new version of E, for the polynomial $\sum_{i=0}^{n-k} b_i z^i$. Now E is an operator on $L_2[0, n-k]$. It easily turns out that the following counterpart of Corollary 3.2 holds.

COROLLARY 3.2^0. *The orthogonal projection $\pi: H^2\to K$, is given by the following*:

(i) $[\pi f](\tau)=f(\tau)$ *for* $\tau\in[0, k)$;

(ii) $[\pi f]^k(\tau)=Q^{-1}\sum_{i=1}^{\infty}[E^{*i}f^{k+i(n-k)}](\tau)$ *for* $\tau\in[0, n-k)$;

(iii) $[\pi f]^{k+i(n-k)}(\tau)=[E^i[\pi f]^k](\tau)$ *for* $\tau\in[0, n-k)$, $\quad i=0, 1, 2, \cdots$.

Of course, Corollaries 3.2 and 3.2^0 are equivalent, but the latter offers a simpler computational tool: the matrices $\tilde{E}$ and $\tilde{Q}$ corresponding to the revised definition of E are of smaller size, $(n-k)\times(n-k)$ instead of $n\times n$, as before. The computation of the coefficients $\tilde{G}_i$ (which will remain $n\times n$) will then require solutions of Lyapunov equations of lower dimensionality, and thus simplify the numerical computations needed for the solution of the maximal eigenvalue/eigenfunction problem, as explained in the following section. For notational simplicity, we shall nonetheless continue the discussion in the general framework of Corollary 3.2 and Observation 3.5.

4. Characterization of eigenvalues and eigenfunctions of T^*T. In what follows we rule out the trivial case $w(s)\equiv\eta$ (for then $T=\eta I$). Our solution strategy is to identify a $2mn$-dimensional subspace $U=U(\lambda^2)\subset X$, associated with each positive number λ^2, such that if λ^2 is an eigenvalue of T^*T, then all its eigenfunctions must belong to U. Then we shall narrow the domain of possible eigenfunctions of λ^2 to a smaller subspace $U_0\subset U$, with dim $U_0=2m$. Finally we shall define an operator $\Omega(\lambda^2)$ on U_0, with the property that $0\neq\xi\in U_0$ is an eigenfunction if and only if $\Omega(\lambda^2)\xi=0$. The operator $\Omega(\lambda^2)$ will have an easy-to-construct matrix representation, so the characterization reduces to simple linear algebra. The original statement of the following useful observation was made by Flamm and Mitter [3], [5], for the particular case $P(z)=z$.

OBSERVATION 4.1. *Suppose that λ^2 is an eigenvalue of T^*T, and that $\xi\in X$ is an associated eigenfunction. Let $U=U(\lambda^2)\subset X$ be the subspace spanned by the functions $\kappa(\tau)$, such that $\tilde{\kappa}(\tau)=\tau^q e^{\mu\tau}\tilde{u}$, where $\tilde{u}$ is a constant n-vector, and where $s=\mu(\neq\infty)$ is a zero of order $\geqq q+1$ of the equation*

$$w^*(s)w(s)=\lambda^2. \tag{4.1}$$

Then $\xi\in U$.

(*Note that* dim $U=2nm$, *unless for the case* $\lambda^2=|\eta|^2$, *where* dim $U=2n(m-1)$.)

Proof. Suppose $\phi=T\xi$ and $\zeta=T^*\phi$, for some ξ, ϕ, and $\zeta\in X$. Denote

$$\tilde{x}_i(\tau)=\alpha_i\left(\int_0^{\tau} e^{\beta_i(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta+\tilde{G}_i\int_0^1 e^{\beta_i(\theta-\tau)}\tilde{\xi}(\theta)\,d\theta\right),$$

$$\tilde{x}_{i+m}(\tau)=\bar{\alpha}_i\left(\int_{\tau}^1 e^{\beta_i(\tau-\theta)}\tilde{\phi}(\theta)\,d\theta+\tilde{Q}^{-1}\tilde{G}_i^*\tilde{Q}\int_0^1 e^{\bar{\beta}_i(\tau-\theta)}\tilde{\phi}(\theta)\,d\theta\right)$$

for $i = 1, 2, \cdots, m$ and $\tau \in [0, 1]$. Then the following equations hold:

$$
\begin{aligned}
&\dot{\tilde{x}}_i = -\beta_i \tilde{x}_i + \alpha_i \tilde{\xi}, \qquad i = 1, \cdots, m, \\
&\tilde{\phi} = \eta \tilde{\xi} + \sum_{i=1}^{m} \tilde{x}_i, \\
&\dot{\tilde{x}}_{i+m} = \bar{\beta}_i \tilde{x}_{i+m} - \bar{\alpha}_i \tilde{\phi}, \qquad i = 1, \cdots, m, \\
&\tilde{\zeta} = \bar{\eta} \tilde{\phi} + \sum_{i=1}^{m} \tilde{x}_{i+m}.
\end{aligned} \tag{4.2}
$$

Equation (4.2) can be written in a standard system form as

$$
\dot{z} = Az + B\tilde{\xi}, \qquad \tilde{\zeta} = Cz + |\eta|^2 \tilde{\xi} \tag{4.2^0}
$$

where z is a $2nm$ vector. We leave out the obvious detail of the matrices A, B, and C, and mention just the following two facts.

The first is captured in the formula

$$
C(s - A)^{-1} B = (W^*(s) W(s) - |\eta|^2) \tilde{I}, \tag{4.3}
$$

where $\tilde{I}$ is the $n \times n$ identity matrix.

The second fact is that for each $k = 0, 1, 2, \cdots$, the matrix $CA^k B$ is a scalar multiple, say, by some $\varepsilon(k)$, of the matrix $\tilde{I}$. In particular, either $CA^k B$ vanishes, or it is an invertible matrix.

Suppose now that λ^2 is an eigenvalue of T^*T associated with the eigenfunction $\xi \in X$. Then $\lambda^2 \tilde{\xi}$ substitutes for $\tilde{\zeta}$ in (4.2^0), and we have the following proposition.

PROPOSITION. *There exists a matrix L such that $\tilde{\xi} = Lz$.*

Proof. The case $\lambda^2 \neq |\eta|^2$ is easier: from the second equation in (4.2^0) we obtain $L = C/(\lambda^2 - |\eta|^2)$.

Suppose that $\lambda^2 = |\eta|^2$. By assumption ($w(s) \not\equiv \eta$, $w(s) = 0 \Rightarrow \operatorname{Re} s < 0$, and $w(s)^{-1} = 0 \Rightarrow \operatorname{Re} s < 0$), there holds $w^*(s) w(s) \not\equiv |\eta|^2$. It thus follows from (4.3) that an integer $k \geqq 0$ exists, such that $CA^j B = 0$ for $0 \leqq j < k$ and $CA^k B \neq 0$. (By Cayley-Hamilton, $k \leqq n - 1$.) As we have noted above, $CA^k B = \varepsilon(k) \tilde{I}$ for some nonzero scalar $\varepsilon(k)$. In particular, $CA^k B$ is an invertible matrix.

Now, $Cz \equiv 0$ since $\lambda^2 = |\eta|^2$ and $\tilde{\zeta} = \lambda^2 \tilde{\xi}$. Thus,

$$
0 \equiv C\dot{z} = CAz + CB\tilde{\xi}.
$$

If $k > 0$ it follows that $CAz \equiv 0$, which in turn implies

$$
0 \equiv CA\dot{z} = CA^2 z + CAB\tilde{\xi}.
$$

Continuing this way inductively, we deduce that $CA^j z \equiv 0$ for $j = 0, 1, \cdots, k$, and consequently that

$$
0 \equiv CA^{k+1} z + CA^k B\tilde{\xi}.
$$

Thus, $L = -(CA^k B)^{-1} CA^{k+1}$ is the desired matrix.

Having established the claim, it follows that $z(\tau)$ satisfies a linear, autonomous, time-invariant ordinary differential equation (ODE) for $\tau \in [0, 1]$. It can therefore be extended analytically to a solution over the positive ray $\tau \geqq 0$. The Laplace transform, $z(s)$, of the extended solution, is a strictly proper rational function. Consequently, $\tilde{\xi}(\tau)$ is analytically extendable for all $\tau \geqq 0$, and the Laplace transform of the extended

function, $\tilde{\xi}(s)$, is a well-defined, strictly proper rational function. Following Laplace transformation, (4.2^0) becomes

$$(\lambda^2-|\eta|^2)\tilde{\xi}(s)=C(s-A)^{-1}(z_0+B\tilde{\xi}(s))$$
$$=C(s-A)^{-1}z_0+(w^*(s)w(s)-|\eta|^2)\tilde{\xi}(s) \quad \text{by (4.3)}\cdots$$

and we finally get

$$\tilde{\xi}(s)=-(w^*(s)w(s)-\lambda^2)^{-1}C(s-A)^{-1}z_0, \tag{4.4}$$

where z_0 is the initial value of the (time domain) function $z(\tau)$.

By definition of A, the poles of $C(s-A)^{-1}z_0$ are poles of $w^*(s)w(s)$. Hence, poles of $\tilde{\xi}(s)$ are finite ($\neq\infty$) zeros of $w^*(s)w(s)-\lambda^2$; that is, they are solutions of (4.1). Hence ξ belongs to U as claimed.

Our next effort is to restrict the initial $2mn$ vector z_0 (from (4.4)) to a subspace of dimension $2m$.

We easily observe that the matrices A, B, and C of (4.2^0) are all built of $n\times n$ blocks which are each a constant multiple of the $n\times n$ identity. Let A_0, B_0, and C_0 be the matrices obtained by collapsing each of these blocks to a scalar entry. Let κ be the $2m\times 2m$ main-diagonal matrix-valued measure on $[0, n]$ where the first m main-diagonal entries are $\alpha_i e^{-\beta_i\tau}d[\sum_{k=0}^{n-1} e^{\beta_i k}a_{ik}\chi_{[k,k+1)}(\tau)]$, $i=1,\cdots,m$ followed by the m entries $\bar{\alpha}_i e^{\beta_i\tau}d[\sum_{k=0}^{n-1} e^{-\bar{\beta}_i k}c_{ik}\chi_{[k,k+1)}(\tau)]$, $i=1,\cdots,m$. Let y be the $2m$-vector whose first m components are $\int_0^1 e^{\beta_i\theta}\langle\tilde{b}_i, \tilde{\xi}(\theta)\rangle\, d\theta$, $i=1,\cdots,m$ followed by $\int_0^1 e^{-\bar{\beta}_i\theta}\langle\tilde{d}_i, \tilde{\phi}(\theta)\rangle\, d\theta$, $i=1,\cdots,m$. Finally, let $x(\tau)$ be the $2m$-vector valued function, whose ith entry, $x_i(\tau)$, is the scalar version of the function $\tilde{x}_i(\cdot)$ from (4.2).

Then, in the same way that (4.2^0) follows from Observation 3.5, Corollary 3.7 implies the dynamics

$$\begin{aligned} dx(\tau)&=[A_0x(\tau)+B_0\xi(\tau)]\,d\tau+d\kappa(\tau)y,\\ \zeta(\tau)&=C_0x(\tau)+|\eta|^2\xi(\tau)\end{aligned} \tag{4.5}$$

for $\tau\in[0, n]$, and with $x(0-)=0$. By arguments similar to those in the proof of Observation 4.1 we deduce that if $\xi(\tau)$ is an eigenfunction for λ^2, then there exists a $1\times 2m$ matrix L_0 with $\xi(\tau)=L_0x(\tau)$.

The first equation in (4.5) thus becomes

$$dx(\tau)=(A_0+B_0L_0)x(\tau)\,d\tau+d\kappa(\tau)y, \tag{4.6}$$

or, in an integrated form (since $x(0-)=0$)

$$x(\tau)=[e^{(A_0+B_0L_0)\cdot}*d\kappa](\tau)y. \tag{4.6^0}$$

Now define functions $x(\tau)$ via (4.6^0) where y is allowed to be any $2m$-vector. The linear mapping from the vector y to the $2mn$ vector $z=(z_{(i-1)n+k}=x_i(k-1+)\colon i=1,\cdots,2m \text{ and } k=1,\cdots,n)$ is therefore a $2mn=2m$ matrix which we denote by M. This matrix depends on λ^2 since L_0 does, and it can be obtained via numerical integration (that is, by computing n different $2m\times 2m$ matrix exponentials).

COROLLARY 4.2. *Fix λ^2 and let $U_0\subset U$ be the space spanned by functions $\xi(\tau)\in L_2[0, n)$, such that $\tilde{\xi}(\tau)$ is the restriction to the interval $\tau\in[0, 1]$ of the inverse Laplace transform of*

$$\tilde{\xi}(s)=(w^*(s)w(s)-\lambda^2)^{-1}C(s-A)^{-1}My$$

for some $2m$-vector y. If λ^2 is an eigenvalue of T^*T, then associated eigenfunctions must belong to U_0.

Proof. From the above, the initial vector $z_0 = z(0)$ in (4.2^0) and (4.4) must be of the form $z_0 = My$ for some $2m$-vector y.

Now, given λ^2, we restrict our attention to the subspace $U_0(\lambda^2)$. For notational convenience we also introduce the subspaces V, W, U_1, V_1, and $W_2 \subset X$ as follows:

$$V = \text{span}\{v(\tau)\colon \tilde{v}(\cdot) = e^{-\beta_i \cdot}\tilde{v} \quad \text{for } i = 1, \cdots, n \text{ and constant } n-\text{vectors } \tilde{v}\},$$

$$W = \text{span}\{w(\tau)\colon \tilde{w}(\cdot) = e^{\bar{\beta}_i \cdot}\tilde{w} \quad \text{for } i = 1, \cdots, n \text{ and constant } m-\text{vectors } \tilde{w}\},$$

$$U_1 = T(U_0) \cap U, \quad V_1 = T(U_0) \cap V, \quad W_2 = T^*(U_1) \cap W.$$

(The latter three spaces depend on λ^2.)

Following from Observation 3.4 the operator T maps U into $U+V$, and T^* maps $U+V$ into $U+V+W$. From the assumptions on $w(s)$ it is born that $U \cap V = \{0\}$, $U \cap W = \{0\}$, and indeed, $V \cap W = \{0\}$. Thus, the appropriate restrictions of T and T^* have the block structures

$$T|_U = \begin{bmatrix} T_1 \\ T_2 \end{bmatrix} \quad \text{and} \quad T^*|_{U+V} = \begin{bmatrix} T_{11}^* & 0 \\ 0 & T_{22}^* \\ T_{31}^* & T_{32}^* \end{bmatrix}$$

where T_1 and $T_{11}^*\colon U \to U$, T_2 and $T_{22}^*\colon V \to V$, $T_{31}^*\colon U \to W$ and $T_{32}^*\colon V \to W$. We denote

$$\Omega(\lambda^2) = \begin{bmatrix} T_2 \\ T_{31}^* T_1 \end{bmatrix}\Bigg|_{U_0}.$$

The following is our main result.

THEOREM 4.3. *The number λ^2 is an eigenvalue of T^*T and is associated with the eigenfunction $\xi(\tau) \in X$ if and only if $\xi(\tau) \in U_0$ and $\Omega(\lambda^2)\xi = 0$.*

Before bringing the proof, let us note that $\Omega(\lambda^2)$ maps U_0 into $V_1 + W_2$. Computed in terms of the formulae in Observation 3.5, a matrix representation of $\Omega(\lambda^2)$ is (at most) $2m \times 2m$, and the theorem becomes a linear algebraic rank condition.

Proof. By Corollary 4.2, λ^2 is an eigenvalue of T^*T associated with the eigenfunction ξ, if and only if $\xi \in U_0$ and the following equations hold:

$$T_{11}^* T_1 \xi = \lambda^2 \xi \tag{4.7}$$

and

$$\begin{bmatrix} 0 & T_{22}^* \\ T_{31}^* & T_{32}^* \end{bmatrix}\begin{bmatrix} T_1 \\ T_2 \end{bmatrix}\xi = 0. \tag{4.8}$$

Notice that (4.8) can be rewritten also as

$$\begin{bmatrix} T_{22}^* & 0 \\ T_{32}^* & I \end{bmatrix}\Omega(\lambda^2)\xi = 0 \tag{4.8^0}$$

so that the theorem will follow from these next two propositions.

PROPOSITION 4.4. $T_{11}^* T_1 = \lambda^2 I_U$.

PROPOSITION 4.5. *T_{22}^* is invertible.*

Proof of Proposition 4.4. It will be easier to establish this claim in the frequency domain, referring to the Laplace transforms of analytic extensions of functions from U. This space is spanned by functions of the form $\tilde{u}/(s-\mu)^q$ where μ is a zero of

order $\geqq q$ of (4.1) and $\tilde{u}$ is a constant n-vector. Now, as in the proof of Observation 4.1, we take the Laplace transform of (4.2^0), and invoke (4.3), to obtain

$$\begin{aligned} T^*T(\tilde{u}/(s-\mu)^q) &= C(S-A)^{-1}(z_0+B\tilde{u}/(s-\mu)^q)+|\eta|^2\tilde{u}/(s-\mu)^q \\ &= C(s-A)^{-1}z_0+w^*(s)w(s)\tilde{u}/(s-\mu)^q \end{aligned} \tag{4.9}$$

where z_0 is a $2mn$ vector. By the assumption on μ and by straightforward factorization of rational functions, the right-hand side of (4.9) is further equal to

$$\cdots = \lambda^2\tilde{u}/(s-\mu)^q + \text{terms with poles at } \{-\beta_i, \bar{\beta}_i, i=1,\cdots,m\},$$

which completes the proof of Proposition 4.4.

Proof of Proposition 4.5. Following from the formula for T^* in Observation 3.5, it turns out that (given a constant vector $\tilde{\xi}$)

$$\tilde{T}^*(\tilde{v}\,e^{-\beta_i\cdot})(t) = w^*(-\beta_i)\tilde{v}\,e^{-\beta_i t} + \text{terms in } e^{\bar{\beta}_i t}, \qquad j=1,\cdots,m.$$

So T_{22}^* has the block diagonal representation $\operatorname{diag}[w^*(-\beta_i)\tilde{I}, i=1,\cdots,n]$. Since $\operatorname{Re}\beta_i>0$ and $w(s)$ is of minimum phase, that matrix is invertible.

Remark. Based on the theorem, the computations needed for the characterization of eigenvalues and eigenfunctions include: (i) linear algebraic computations (including a set of Lyapunov matrix equations that are solved once, at a preparatory stage): (ii) solution of (4.1) (for various λ^2), which is a polynomial equation; (iii) numerical integration, in order to compute the matrices $M(\lambda^2)$. We can avoid the need for numerical integrations, at a cost. The theorem remains true (with the same proof), if we extend the definition of $\Omega(\lambda^2)$ to the whole of U, thus relying on Observation 4.1 instead of on Corollary 4.2. The cost is, of course, that now the matrix representation of $\Omega(\lambda^2)$ will be of a larger size, $2nm\times 2nm$ ($2nm\times 2n(m-1)$, if $\lambda^2=|\eta|^2$) instead of $2m\times 2m$.

Indeed, the latter has been the scheme suggested in an earlier version of this article. The current form of the theorem is based on Observation 3.6, and is motivated by the observation that the rank of the operator $(I-\pi)(e^{-\beta\cdot}*\cdot)|_K$ is at most one.

We conclude this section, as promised, with the following proof.

Proof of Observation 2.2. To be able to apply Weyl's lemma, which we need in the proof, it should be first checked that $T^*T-|\eta|^2$ is a compact operator. And indeed it is: By the formulae given in Observation 3.5, both $T-\eta$ and $T^*-\bar{\eta}$ are of the form of Volterra plus finite-dimensional operators on $L_2([0,1],\mathbf{C}^n)$, and the latter space is homeomorphic to X. Hence $T-\eta$ and $T^*-\bar{\eta}$ are compact, and so is $T^*T-|\eta|^2=(T^*-\bar{\eta})(T-\eta)+\bar{\eta}(T-\eta)+\eta(T^*-\bar{\eta})$. (A more general, yet abstract, argument was suggested by the referee. The compactness of $T^*T-|\eta|^2$ follows from the continuity of $G^*(s)w(s)$ on the extended imaginary axis, via a Hartman–Theorem-type result of Muhly [19].)

Having made this observation, we deduce that the norm $\|T^*T\|$ equals the spectral radius $\rho(T^*T)$ and that $|\eta|^2$ is the only limit point of eigenvalues of T^*T. We shall now prove Observation 2.2, point by point.

(ii) Suppose λ^2 were the maximal eigenvalue of T^*T. By Weyl's Lemma $\rho(T^*T)=\|T^*T\|=\|T\|^2$ and $|\eta|^2\leqq\lambda^2$. Obviously, $\|T\|\leqq\|w\|_\infty$. Hence $\lambda^2\in[|\eta|^2,\|w\|_\infty^2]$.

(iii) Suppose $\|w\|_\infty=|\eta|$. We then have the "sandwich" argument $|\eta|^2\leqq\rho(T^*T)=\|T^*T\|=\|T\|^2\leqq\|w\|_\infty^2=|\eta|^2$, which implies that $\|T\|=\|w\|_\infty$. Hence $w(s)$ interpolates T.

Yet if $w(s)\not\equiv\eta$ then $|\eta|^2$ cannot be an eigenvalue of T^*T. Suppose it were; then Sarason's Lemma 2.1 tells us that $w(s)$ should be all-pass, that is, a constant multiple of a Blascke product. In particular then, $w(s)$ cannot be of minimum phase, as assumed.

(i) Since $|\eta|^2$ is the only limit point of eigenvalues of T^*T, it suffices to verify the existence of one eigenvalue $\lambda^2 > |\eta|^2$ to deduce the existence of a maximal eigenvalue. The equality $\rho(T^*T) = \|T^*T\|$ further reduces the proof to a search for a function $f \in K$ such that $\|Tf\|_2^2 > |\eta|^2 \|f\|_2^2$. We now establish the existence of such a function.

We shall use the following fact which is stated in the frequency domain. (Namely, if $f(s)$ is a member of K, we restrict our attention to its values over the imaginary axis, $s = j\omega$.)

PROPOSITION. *The restrictions of functions from K to any compact interval $[-j\omega_1, j\omega_1]$ form a dense subspace of $L_2[-j\omega_1, j\omega_1]$.*

Proof. Following from Corollary 3.3, a frequency domain description of K is given by

$$K = \left\{ f(j\omega) = \zeta(j\omega) \int_0^1 e^{-j\omega\tau} \tilde{\xi}(\tau)\, d\tau \colon \tilde{\xi} \in L_2[0, 1] \right\}$$

where

$$\zeta(j\omega) = [1, e^{-j\omega}, \cdots, e^{-(n-1)j\omega}](\tilde{I} - \tilde{E}\, e^{-j\omega n})^{-1}.$$

Suppose that $g(j\omega) \in L_2[-j\omega_1, j\omega_1]$ were orthogonal to all members of K, over the said interval. Then

$$\begin{aligned} 0 &= \int_{-j\omega_1}^{j\omega_1} \left\langle g(j\omega), \zeta(j\omega) \int_0^1 e^{-j\omega\tau} \tilde{\xi}(\tau)\, d\tau \right\rangle d\omega \\ &= \int_0^1 \left\langle \int_{-j\omega_1}^{j\omega_1} e^{j\omega\tau} \zeta^*(j\omega) g(j\omega)\, d\omega, \tilde{\xi}(\tau) \right\rangle d\tau \end{aligned}$$

for all $\tilde{\xi}(\tau) \in L_2[0, 1]$. Hence

$$\int_{-j\omega_1}^{j\omega_1} e^{j\omega\tau} \zeta^*(j\omega) g(j\omega)\, d\omega = 0 \quad \text{for } \tau \in [0, 1],$$

and by Plancherals equality and the analyticity of the Fourier transform, it follows that $\zeta^*(j\omega) g(j\omega) \equiv 0$. Consequently, $g = 0$, as claimed.

Denote by $R : K \to K^\perp$ the mapping $f \to (I - \pi)(w * f)$. Following from the proof of Observation 3.6, rank $R \leqq m$.

Recall that by our assumption in part (i) of Observation 2.2, there exists some $\omega_0 > 0$ such that $|\omega| > |\omega_1|$ implies $|w(j\omega)| > |\eta|$. Thus there exist $\varepsilon > 0$ and $\omega_2 > \omega_1 > \omega_0$, such that for $|\omega| \in [\omega_1, \omega_2]$ there holds $|w(j\omega)| > |\eta| + \varepsilon$. Given ω_1 and ω_2, let $Y \subset L_2[-j\omega_2, j\omega_2]$ be the space formed by the L_2 closure of restrictions of functions from $\ker(R)$ to the interval $[-j\omega_2, j\omega_2]$. Let $Z \subset Y$ be the subspace of functions from Y that vanish along $[-j\omega_1, j\omega_1]$. Following from the proposition above, codim $Y < \infty$ and dim $Z = \infty$.

Now choose $\delta > 0$ such that $(|\eta| + \varepsilon)^2(1 - \delta) > (|\eta| + \varepsilon/2)^2$. Every nonzero function $g \in Z$ can be approximated (in $L_2[-j\omega_2, j\omega_2]$) by the restriction to $[-j\omega_2, j\omega_2]$ of some function $f \in \ker(R)$, which satisfies

$$\int_{|\omega| \in [\omega_1, \omega_2]} |f(j\omega)|^2\, d\omega > (1 - \delta) \int_{-\omega_2}^{\omega_2} |f(j\omega)|^2\, d\omega.$$

In particular, we conclude that there exists at least one function $f \in \ker(R)$ that satisfies this inequality.

Since $f \in \ker(R)$, there holds $Tf(s) = w(s)f(s)$, whence,

$$\begin{aligned}
\|Tf\|_2^2 &= \int_{-\infty}^{\infty} |w(j\omega)f(j\omega)|^2 \, d\omega \\
&\geqq \left(\int_{|\omega|>\omega_2} + \int_{|\omega|\in[\omega_1,\omega_2]} \right) |w(j\omega)f(j\omega)|^2 \, d\omega \\
&\geqq |\eta|^2 \int_{|\omega|>\omega_2} |f(j\omega)|^2 \, d\omega + (|\eta|+\varepsilon)^2 \int_{|\omega|\in[\omega_1,\omega_2]} |f(j\omega)|^2 \, d\omega \\
&\geqq \left(|\eta|^2 \int_{|\omega|>\omega_2} + (|\eta|+\varepsilon)^2(1-\delta) \int_{|\omega|\leqq\omega_2} \right) |f(j\omega)|^2 \, d\omega \\
&\geqq \left(|\eta|^2 \int_{|\omega|>\omega_2} + \left(|\eta|+\frac{\varepsilon}{2}\right)^2 \int_{|\omega|\leqq\omega^2} \right) |f(j\omega)|^2 \, d\omega \\
&> |\eta|^2 \|f\|_2^2.
\end{aligned}$$

As indicated above, the existence of f completes the proof of part (i), and hence of the observation.

5. A simple example. Set $P(z) = 1 - e^{\alpha} z$ and $w(s) = 1/(\beta + s)$ where α and β are real, positive numbers and $\alpha \neq \beta$. Then $n = 1$, and the matrices $\tilde{E}$ and $\tilde{Q}$ are scalars. We get $\tilde{E} = e^{-\alpha}$ and $\tilde{Q} = 1/(1 - e^{-2\alpha})$. Computation of the coefficient $\tilde{G} = \tilde{G}_1$ yields $\tilde{G} = e^{-(\alpha+\beta)}/(1 - e^{-(\alpha+\beta)})$.

In our example, (4.1) takes the form

$$\frac{1}{\beta + s} \cdot \frac{1}{\beta - s} = \lambda^2,$$

and solutions satisfy $s = \pm\varepsilon$, for $\varepsilon = (\sqrt{\beta^2\lambda^2 - 1})/\lambda$. So (given λ), the space $U = U_0$ is spanned by $\xi_1(\tau) = e^{\varepsilon\tau}$ and $\xi_2(\tau) = e^{-\varepsilon\tau}$. Computations of the matrices T_1, T_2, and T_{31}^* yield

$$T_1 = \begin{bmatrix} \dfrac{1}{\beta+\varepsilon} & 0 \\ 0 & \dfrac{1}{\beta-\varepsilon} \end{bmatrix},$$

$$T_2 = \frac{1}{1 - e^{-(\alpha+\beta)}} \left[\frac{e^{\varepsilon-\alpha} - 1}{\beta+\varepsilon}, \frac{e^{-\varepsilon-\alpha} - 1}{\beta-\varepsilon} \right],$$

$$T_{31}^* = \frac{e^{-(\alpha+\beta)}}{1 - e^{-(\alpha+\beta)}} \left[\frac{e^{\varepsilon+\alpha} - 1}{\beta-\varepsilon}, \frac{e^{-\varepsilon+\alpha} - 1}{\beta+\varepsilon} \right].$$

So $\Omega(\lambda^2)$ is given by

$$\Omega(\lambda^2) = \frac{1}{1 - e^{-(\alpha+\beta)}} \begin{bmatrix} \dfrac{e^{\varepsilon-\alpha} - 1}{\beta+\varepsilon} & \dfrac{e^{-\varepsilon-\alpha} - 1}{\beta-\varepsilon} \\ \dfrac{(e^{\varepsilon+\alpha} - 1)\, e^{-(\alpha+\beta)}}{(\beta-\varepsilon)(\beta+\varepsilon)} & \dfrac{(e^{-\varepsilon+\alpha} - 1)\, e^{-(\alpha+\beta)}}{(\beta+\varepsilon)(\beta-\varepsilon)} \end{bmatrix}$$

$$= \frac{1}{1 - e^{-(\alpha+\beta)}} \begin{bmatrix} 1 & 0 \\ 0 & e^{-(\alpha+\beta)} \end{bmatrix} \cdot \begin{bmatrix} e^{\varepsilon-\alpha} - 1 & e^{-\varepsilon-\alpha} - 1 \\ e^{\varepsilon+\alpha} - 1 & e^{-e+\alpha} - 1 \\ \hline \beta-\varepsilon & \beta+\varepsilon \end{bmatrix} \cdot \begin{bmatrix} \dfrac{1}{\beta+\varepsilon} & 0 \\ 0 & \dfrac{1}{\beta-\varepsilon} \end{bmatrix}.$$

The rank condition (since $\Omega(\lambda^2)$ is 2×2) is that λ^2 is an eigenvalue when

$$(\beta-\varepsilon)(e^{\varepsilon-\alpha}-1)(e^{-\varepsilon+\alpha}-1)-(\beta+\varepsilon)(e^{-\varepsilon-\alpha}-1)(e^{\varepsilon+\alpha}-1)=0. \tag{5.1}$$

In this example we have $\|w(s)\|_\infty = 1/\beta$ and $\eta = 0$. By Observation 2.2 a maximal eigenvalue exists, and it belongs to the interval $(0, 1/\beta^2]$. When λ^2 descends from $1/\beta^2$ to zero, the corresponding ε ranges from zero to $+j\infty$ along the imaginary axis. Thus, our scheme prompts a search for the minimal imaginary solution $\varepsilon = j\omega$ to (5.1). Given such a solution, and the corresponding eigenvalue λ^2, it is easy to find a zero vector $x=[x_1, x_2]'$ for $\Omega(\lambda^2)$. Given such x, a maximal function for T in the space X would be

$$\xi(\tau) = x_1 e^{\varepsilon\tau} + x_2 e^{-\varepsilon\tau}.$$

To use Sarason's Lemma, we also need $T\xi(\tau)$, which is

$$T\xi(\tau) = \frac{x_1}{\beta+\varepsilon} e^{\varepsilon\tau} + \frac{x_2}{\beta-\varepsilon} e^{-\varepsilon\tau} + \frac{1}{1-e^{-(\alpha+\beta)}}\left[\frac{x_1(e^{\varepsilon-\alpha}-1)}{\beta+\varepsilon} + \frac{x_2(e^{-\varepsilon-\alpha}-1)}{\beta-\varepsilon}\right] e^{-\beta\tau}.$$

Using Corollary 3.2 (as in the proof of Observation 2.2, above), we obtain the Laplace transforms of the corresponding members of K:

$$[\xi](s) = \frac{1}{1-e^{-(\alpha+s)}}\left[\frac{x_1(e^{\varepsilon-s}-1)}{\varepsilon-s} + \frac{x_2(e^{-\varepsilon-s}-1)}{-\varepsilon-s}\right],$$

$$[T\xi](s) = \frac{1}{1-e^{-(\alpha+s)}}\left[\frac{x_1(e^{\varepsilon-s}-1)}{(\beta+\varepsilon)(\varepsilon-s)} + \frac{x_2(e^{-\varepsilon-s}-1)}{(\beta-\varepsilon)(-\varepsilon-s)} + \frac{e^{-(\alpha+\beta)}(e^{-\beta-s}-1)}{(1-e^{-(\alpha+\beta)})(-\beta-s)}\left[\frac{x_1(e^{\varepsilon-\alpha}-1)}{\beta+\varepsilon} + \frac{x_2(e^{-\varepsilon-\alpha}-1)}{\beta-\varepsilon}\right]\right].$$

The interpolating function would then be $[T\xi](s)/[\xi](s)$.

REFERENCES

[1] T. E. DJAFERIS AND S. K. MITTER, *Algebraic methods for the study of some linear matrix equations*, Linear Algebra Appl., 44 (1982), pp. 125-142.

[2] F. FAGNANI, D. S. FLAMM, AND S. K. MITTER, *Some min-max optimization problems in infinite dimensional control systems*, LIDS Technical Report P-1640, January 1987.

[3] D. S. FLAMM, *Control of delay systems for Minimax sensitivity*, Ph.D. Thesis, LIDS-TH 1560 (June 1986), Massachusetts Institute of Technology, Cambridge, MA.

[4] D. S. FLAMM AND S. K. MITTER, *Progress on H^∞ optimal sensitivity for delay systems* I: *minimum phase plant with input delay, 1 pole/zero weighting function*, LIDS Technical Report P-1513, Massachusetts Institute of Technology, Cambridge, MA, November 1985.

[5] ———, *Progress on H^∞ optimal sensitivity for delay systems* II: *minimum phase plant with input delay, general rational weighting function*, preprint, LIDS, Massachusetts Institute of Technology, Cambridge, MA, January 1986.

[6] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *Weighted sensitivity minimization for delay systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 763-766.

[7] C. FOIAS AND A. TANNENBAUM, *On the H^∞-optimal sensitivity problem for systems with delays*, SIAM J. Control Optim., to appear.

[8] ———, *On the Nehari problem for a certain class of L^∞ functions appearing in control theory*, J. Funct. Anal., to appear.

[9] ———, *On the uniqueness of a minimal norm representative of an operator in the commutant of the compressed shift*, in Proc. of Amer. Math. Soc., to appear.

[10] C. FOIAS, A. TANNENBAUM, AND G. ZAMES, *Sensitivity minimization for arbitrary* SISO *distributed plants*, Systems Control Lett., 8 (1987), pp. 189-195.

[11] ———, *Some explicit formulae for the singular values of certain Hankel operators with factorizable symbol*, SIAM J. Math. Anal., to appear.

[12] B. A. FRANCIS, *A Course in H^∞ Control Theory*, Lecture Notes in Control and Information Sciences, 88, Springer-Verlag, Berlin, 1987.
[13] B. A. FRANCIS AND J. DOYLE, *Linear control theory with an H^∞ optimality criterion*, Systems Control Group Report #8501, University of Toronto, Ontario, Canada, October 1985.
[14] J. B. GARNETT, *Bounded Analytical Functions*, Academic Press, New York, 1981.
[15] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.
[16] P. R. HALMOS, *A Hilbert Space Problem Book*, American Book Co., New York, 1967.
[17] J. W. HELTON, *Worst case analysis in the frequency domain: the H^∞ approach to control*, IEEE Trans. Automat. Control, 30 (1985), pp. 1154–1170.
[18] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realization and approximation of linear infinite dimensional systems with error bounds*, CUED/F-CAMS/TR 258, Cambridge University, UK, 1986.
[19] P. MUHLY, *Compact operators in the commutant of a contraction*, J. Funct. Anal., 8 (1971), pp. 197–224.
[20] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179–203.
[21] G. TADMOR, *An interpolation problem associated with H^∞ optimal design in systems with distributed input lags*, Systems Control Lett., 8 (1987), pp. 313–319.
[22] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, seminorms and approximate inverses*, IEEE Trans. Automat. Control, 23 (1981), pp. 301–320.
[23] G. ZAMES AND B. A. FRANCIS, *Feedback minimax sensitivity and optimal robustness*, IEEE Trans. Automat. Control, 28 (1983), pp. 585–601.
[24] G. ZAMES, A. TANNENBAUM, AND C. FOIAS, *Optimal H^∞ interpolation: a new approach*, in Proc. Conference on Decision and Control, 1986, pp. 350–355.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 527-545, May 1989

005

ADMISSIBILITY OF UNBOUNDED CONTROL OPERATORS*

GEORGE WEISS†

Abstract. For linear systems described by $\dot{x}(t) = Ax(t) + Bu(t)$, where A generates a semigroup on the state space X and B is an unbounded operator, some necessary as well as some sufficient conditions are given for B to be admissible, i.e., for any t, the state $x(t)$ should be in X and should depend continuously on the input $u \in L^p$. This approach begins with an axiomatic description of such a system in terms of a functional equation. The results are applied to the wave equation on a bounded interval.

Key words. infinite-dimensional linear systems, unbounded control operators, admissibility, multipliers

AMS(MOS) subject classifications. 93C25, 93C20

1. Introduction. In this paper we deal with infinite-dimensional linear time-invariant systems. If we denote by $x(t)$ the state at time t of such a system, then the evolution of $x(t)$ is, in a certain sense, described by the differential equation

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

Here $x(t) \in X$, the Banach space X being the state space, $u(t) \in U$, the Banach space U being the input space, and A is the generator of a strongly continuous semigroup $\mathbb{T}$ on X. The input function $u(\cdot)$ is assumed to be locally L^p for some $p \in [1, \infty]$.

Our interest will focus on the linear operator B, the *control operator* of the system. B is called *bounded* if it is a bounded operator from U to X, and *unbounded* if it is a bounded operator from U to some larger Banach space V,

$$X \subset V,$$

but not from U to X. (This terminology may seem strange but it is now generally agreed upon.)

Unbounded control operators appear naturally, for example, when we model boundary or point control for systems described by linear PDE's. There is extensive literature dealing with systems having unbounded control operators, for example, Curtain and Pritchard [4, Chap. 8], Curtain and Salamon [5], Desch, Lasiecka, and Schappacher [7], Ho and Russell [12], Lasiecka [14], Lasiecka and Triggiani [15]-[17], Pritchard and Wirth [23], Russell [25], Salamon [26]-[29], and Washburn [30].

Here we address the following problems: which operators B should be accepted as "legitimate," and how to recognize them. We want to make those points somewhat more precise.

Assume that X is dense in V and the semigroup $\mathbb{T}$ has a continuous extension to V (denoted by the same symbol). This assumption might seem artificial at this stage but we shall see later that it is a consequence of natural assumptions about the system we want to model by (1.1). By a solution $x(\cdot)$ of (1.1), for initial conditions given at, say, $t = 0$, we mean the function defined for $t \geqq 0$ by the variation of parameters formula

$$x(t) = \mathbb{T}_t x(0) + \int_0^t \mathbb{T}_{t-\sigma} Bu(\sigma)\, d\sigma. \tag{1.2}$$

* Received by the editors May 6, 1987; accepted for publication (in revised form) May 31, 1988.
† Department of Theoretical Mathematics, Weizmann Institute, Rehovot 76100, Israel.

For the formula above to define an X-valued function, the integral in (1.2) must be in the state space X, despite the fact that what we integrate is in V. If that is indeed the case, for any $t \geqq 0$ and any p-integrable U-valued function $u(\cdot)$ on $[0, t]$, then we say that for the given semigroup $\mathbb{T}$, the control operator B is *admissible*.

Admissible control operators that yield (through (1.2)) the same solution $x(\cdot)$, for any given $x(0)$ and $u(\cdot)$, will be identified.

We shall show that, given the state space X and the semigroup $\mathbb{T}$ on X, there is a Banach space X_{-1}, the *same for all* admissible control operators B (for various U and p) such that X is dense in X_{-1}, $\mathbb{T}$ has a continuous extension to X_{-1}, and B is a bounded operator from U to X_{-1} (see § 3).

However, even for fixed U and p, there is generally no Banach space V with the property that B is admissible if and only if it belongs to $\mathscr{L}(U, V)$ (the space of bounded operators from U to V) (see § 5).

Example 1.1. Consider the wave equation on $[0, \pi]$, without input for the time being:

$$\frac{\partial^2}{\partial t^2}\psi(\zeta, t) = \frac{\partial^2}{\partial \zeta^2}\psi(\zeta, t), \qquad \psi(0, t) = \psi(\pi, t) = 0,$$

$$\psi(\zeta, 0) = \psi_0(\zeta), \qquad \frac{\partial}{\partial t}\psi(\zeta, 0) = \psi_1(\zeta).$$

The weak solution ψ, as a function of ζ, is supposed to be absolutely continuous and to have its ζ-derivative in $L^r[0, \pi]$, for a given $r \in [1, \infty)$. For that, the initial data ψ_0 and ψ_1 have to be given accordingly.

To translate these equations into the semigroup language, let us denote

$$W_0^{1,r}[0, \pi] = \left\{ x \in \mathrm{AC}\,[0, \pi] \,\middle|\, \frac{d}{d\zeta} x \in L^r[0, \pi], x(0) = x(\pi) = 0 \right\},$$

$$W^{2,r}[0, \pi] = \left\{ x \in \mathrm{AC}\,[0, \pi] \,\middle|\, \frac{d}{d\zeta} x \in \mathrm{AC}\,[0, \pi], \frac{d^2}{d\zeta^2} x \in L^r[0, \pi] \right\};$$

AC means "absolutely continuous." We consider these spaces equipped with their usual L^r-type norms. Introducing the state variable and state space

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} \psi(\cdot, t) \\ \dot{\psi}(\cdot, t) \end{pmatrix}, \qquad X = \begin{matrix} W_0^{1,r} \\ \times \\ L^r \end{matrix},$$

we have

$$\dot{x}(t) = Ax(t),$$

where

$$A = \begin{pmatrix} 0 & I \\ d^2/d\zeta^2 & 0 \end{pmatrix}, \qquad D(A) = \begin{matrix} W^{2,r} \cap W_0^{1,r} \\ \times \\ W_0^{1,r} \end{matrix},$$

and A generates a contraction group $\mathbb{T}$ on X.

We take the simplest input space, $U = \mathbb{R}$, and $p \in [1, \infty]$. A bounded control operator B may be identified with an element $b \in X$, having components b_1 and b_2.

The corresponding partial differential equation is

$$\frac{\partial^2}{\partial t^2}\psi(\zeta, t)=\frac{\partial^2}{\partial \zeta^2}\psi(\zeta, \tau)+b_1(\zeta)\frac{\partial}{\partial t}u(t)+b_2(\zeta)u(t).$$

Concerning unbounded control operators for the above semigroup, we shall prove the following. The admissible control operators B are the image through a certain isomorphism of those periodic distributions on $[0, 2\pi]$ whose Fourier coefficients are in the space of multipliers (L^p, L^r). Like the bounded control operators, the unbounded ones have two components b_1 and b_2 that are both distributions on $[0, \pi]$, so the PDE above is satisfied in a certain distributional sense.

We outline the contents of the following sections. In §§ 2 and 3 we derive a general necessary condition for B to be admissible (in fact, we do more than that). Our approach is the following. Suppose B is admissible and define for $t \geqq 0$ and $u \in L^p([0, \infty), U)$

$$\Phi_t u=\int_0^t \mathbb{T}_{t-\sigma}Bu(\sigma)\,d\sigma. \tag{1.3}$$

Then $\Phi=(\Phi_t)_{t\geqq 0}$ is a family of bounded linear operators from $L^p([0, \infty), U)$ to X. (Clearly Φ_t depends only on the restriction of u to $[0, t]$ but we want to avoid the unnecessary complications that would arise if the domain of Φ_t depended on t.)

The semigroup $\mathbb{T}$ and the family Φ satisfy a natural functional equation (see (2.1)) called the *composition property* (Kalman, Falb, and Arbib [13, p. 6]). We define an *abstract linear control system* as a pair $(\mathbb{T}, \Phi)$, where $\mathbb{T}$ is a strongly continuous semigroup and Φ is a family of operators such that the composition property holds. Clearly (1.1), if B is admissible, defines an abstract linear control system via (1.3). The latter is a simple and natural concept in whose definition no mention of unbounded operators is needed. In § 2 we derive some properties of such systems which will be needed later.

In § 3 we prove a representation theorem stating that for $p<\infty$ any abstract linear control system is described by (1.1), with B admissible, and moreover $B \in \mathscr{L}(U, X_{-1})$, where X_{-1} is a certain extension of X depending only on $\mathbb{T}$. The necessary condition $B \in \mathscr{L}(U, X_{-1})$ may be regarded as an upper limit for "how unbounded" an admissible B may be.

In § 4 we introduce the Banach space of admissible control operators B for given U, X, $\mathbb{T}$, and p, denoted $\mathscr{B}_p$, determine $\mathscr{B}_1$ for reflexive X, and give necessary and sufficient conditions for $B \in \mathscr{B}_p$ in the case of invertible semigroups.

In § 5 we first deal with the periodic left shift semigroup on $L^r[0, 2\pi]$, describing $\mathscr{B}_p$ in terms of multipliers. Then we turn to the controlled wave equation of Example 1.1 which is closely related to the controlled periodic left shift.

Some of the abstract machinery developed in §§ 2–4 of this paper has already been used (see Weiss [31]). For admissible unbounded observation operators, see Weiss [32].

2. Abstract linear control systems. We begin by giving the formal definition of an abstract linear control system, as announced in § 1. For that we need the notion of *concatenation* on $\Omega=L^p([0, \infty), U)$, where U is a Banach space.

Let $u, v \in \Omega$ and let $\tau \geqq 0$. Then the *τ-concatenation of u and v*, $u \underset{\tau}{\diamond} v \in \Omega$, is given by

$$\left(u \underset{\tau}{\diamond} v\right)(t)=\begin{cases} u(t) & \text{for } t \in [0, \tau), \\ v(t-\tau) & \text{for } t \geqq \tau. \end{cases}$$

Recall that we work with Ω because we want to define our system as receiving U-valued locally p-integrable input functions, and any segment of such an input function can be thought of as the restriction to a bounded interval of an element of Ω.

DEFINITION 2.1. Let U and X be Banach spaces, $p \in [1, \infty]$ and $\Omega = L^p([0, \infty), U)$. An *abstract linear control system* on X and Ω is a pair

$$\Sigma = (\mathbb{T}, \Phi),$$

where $\mathbb{T} = (\mathbb{T}_t)_{t \geqq 0}$ is a strongly continuous semigroup on X and $\Phi = (\Phi_t)_{t \geqq 0}$ is a family of bounded operators from Ω to X such that

$$\Phi_{\tau+t}\left(u \underset{\tau}{\diamond} v\right) = \mathbb{T}_t \Phi_\tau u + \Phi_t v \tag{2.1}$$

for any $u, v \in \Omega$ and any $\tau, t \geqq 0$.

The functional equation (2.1) is called the *composition property*. The operators Φ_t are called *input maps*.

Remark 2.2. Taking $\tau = t = 0$ in (2.1) we get $\Phi_0 = 0$, whence, taking now only $t = 0$ in (2.1), we get that Φ is *causal*, i.e., for any $\tau \geqq 0$

$$\Phi_\tau = \Phi_\tau P_\tau, \tag{2.2}$$

where P_τ is the projection of Ω onto $L^p([0, \tau], U)$ defined by

$$P_\tau u = u \underset{\tau}{\diamond} 0.$$

In practice, an abstract linear control system can be given in the form (1.1) and (1.2) (the "semigroup formulation") as in Curtain and Pritchard [4], or as a boundary control system (i.e., described by a nonhomogeneous boundary value problem), as in Lions and Magenes [20], Russell [25], or as a neutral functional differential equation, as in Salamon [26]. The system can also be modeled using a strongly continuous cosine operator, as in Lasiecka and Triggiani [15]. The problem of translation of the boundary value formulation into the semigroup formulation is discussed in Curtain [3], Curtain and Salamon [5], Desch, Lasiecka, and Schappacher [7], Fattorini [11], and Washburn [30]. For an abstract and very general treatment of when and how this translation is possible, see Salamon [27] or [28].

PROPOSITION 2.3. *Let X and Ω be as in Definition* 2.1 *with $p < \infty$ and let $\Sigma = (\mathbb{T}, \Phi)$ be an abstract linear control system on X and Ω. Then the function*

$$\varphi(t, u) = \Phi_t u$$

is continuous on the product $[0, \infty) \times \Omega$, in particular $\Phi = (\Phi_t)_{t \geqq 0}$ is a strongly continuous family of operators.

Proof. Taking in (2.1) $u = 0$ and taking the supremum for $\|v\| = 1$ we get, denoting $T = \tau + t$,

$$\|\Phi_t\| \leqq \|\Phi_T\| \quad \text{for} \quad t \leqq T, \tag{2.3}$$

i.e., $\|\Phi_t\|$ is nondecreasing.

Let us first prove the continuity of $\varphi(t, u)$ with respect to the time t, so for the time being let $u \in \Omega$ be fixed and let

$$f(t) = \Phi_t u.$$

Inequality (2.3) together with causality (2.2) implies that for $t \in [0, 1]$

$$\|f(t)\| \leqq \|\Phi_1\| \cdot \|P_t u\|.$$

Obviously $\|P_t u\| \to 0$ for $t \to 0$ (because of $p < \infty$), so

$$\lim_{t \to 0} f(t) = 0.$$

The right continuity of f in any $\tau > 0$ now follows easily from the composition property (2.1). To prove the left continuity of f in $\tau > 0$ we take a sequence (ε_n) with $\varepsilon_n \in [0, \tau]$ and $\varepsilon_n \to 0$ and define $u_n(t) = u(\varepsilon_n + t)$, so $u_n \in \Omega$ and $u_n \to u$ (because of $p < \infty$). We have

$$u = u \underset{\varepsilon_n}{\diamond} u_n,$$

so according to (2.1)

$$\Phi_{\varepsilon_n + (\tau - \varepsilon_n)} u = \mathbb{T}_{\tau - \varepsilon_n} \Phi_{\varepsilon_n} u + \Phi_{\tau - \varepsilon_n} u_n.$$

From here

$$\Phi_\tau u - \Phi_{\tau - \varepsilon_n} u = \mathbb{T}_{\tau - \varepsilon_n} \Phi_{\varepsilon_n} u + \Phi_{\tau - \varepsilon_n}(u_n - u),$$

which yields

$$\|\Phi_\tau u - \Phi_{\tau - \varepsilon_n} u\| \leqq M \cdot \|f(\varepsilon_n)\| + \|\Phi_\tau\| \cdot \|u_n - u\|,$$

where M is a bound for $\|\mathbb{T}_t\|$ on $[0, \tau]$. Thus the left continuity of f in any $\tau > 0$ is also proved.

The joint continuity of φ follows easily now from the decomposition

$$\Phi_t v - \Phi_\tau u = \Phi_t(v - u) + (\Phi_t - \Phi_\tau) u$$

where $(t, v) \to (\tau, u)$. □

PROBLEM 2.4. In the above proof we have twice used the fact that $p < \infty$. I do not know if the proposition holds for $p = \infty$.

We give an estimate for the growth rate of $\|\Phi_t\|$.

PROPOSITION 2.5. *Let X and Ω be as in Definition* 2.1 *and let $(\mathbb{T}, \Phi)$ be an abstract linear control system on X and Ω.*

If $M \geqq 1$ and $\omega > 0$ are such that

$$\|\mathbb{T}_t\| \leqq M e^{\omega t} \quad \forall t \geqq 0,$$

then there is some $L \geqq 0$ such that

$$\|\Phi_t\| \leqq L e^{\omega t} \quad \forall t \geqq 0.$$

Proof. From (2.1) we get through induction

$$\Phi_n\left(\cdots\left(u_1 \underset{1}{\diamond} u_2\right) \underset{2}{\diamond} \cdots \underset{n-1}{\diamond} u_n\right) = \mathbb{T}_{n-1}\Phi_1 u_1 + \mathbb{T}_{n-2}\Phi_1 u_2 + \cdots + \Phi_1 u_n, \tag{2.4}$$

whence for $t \in (n-1, n]$ (using (2.3))

$$\|\Phi_t\| \leqq \|\Phi_n\| \leqq (\|\mathbb{T}_{n-1}\| + \|T_{n-2}\| + \cdots + \|I\|) \cdot \|\Phi_1\|$$

$$\leqq M \frac{e^{\omega n} - 1}{e^\omega - 1} \|\Phi_1\|,$$

so we can take

$$L = M \frac{e^\omega}{e^\omega - 1} \|\Phi_1\|.$$

□

Remark 2.6. For $\omega = 0$, Φ does not have to be uniformly bounded but we can easily obtain from (2.4), using the Cauchy-Hölder inequality, that

$$\|\Phi_t\| \leqq L \cdot (1+t)^{1-1/p} \quad \forall t \geqq 0,$$

for some $L \geqq 0$. For $\omega < 0$, Φ is uniformly bounded.

Remark 2.7. Let $\tilde{\Omega} = L^p_{\text{loc}}([0, \infty), U)$. Concatenation and the projections P_τ have obvious extensions to $\tilde{\Omega}$. $\tilde{\Omega}$ is a Fréchet space with the family of seminorms $p_n(u) = \|P_n u\|$, $n \in \mathbb{N}$. Any family of input maps Φ defined on Ω can be extended to $\tilde{\Omega}$ by continuity, which is the same as using formula (2.2) as a definition.

3. The representation theorem.

DEFINITION 3.1. Let X be a Banach space and $\mathbb{T}$ a strongly continuous semigroup on X with generator $A: D(A) \to X$. Let $\beta \in \rho(A)$, the resolvent set of A (if X is real, take $\beta \in \mathbb{R}$). We define the space X_1 to be $D(A)$ with the norm

$$\|x\|_1 = \|(\beta I - A)x\|,$$

and the space X_{-1} to be the completion of X with respect to the norm

$$\|x\|_{-1} = \|(\beta I - A)^{-1} x\|.$$

Remark 3.2. It is easy to verify that for any different $\beta_1 \in \rho(A)$ instead of β we get equivalent norms $\|\cdot\|_1$ and $\|\cdot\|_{-1}$ (so X_{-1} does not depend on β). In particular, $\|\cdot\|_1$ is equivalent to the graph norm on $D(A)$, so X_1 is complete.

The spaces X_1, X_{-1} appear, for example, in Nagel [21, p. 19] and Da Prato [6]. If X is *reflexive* then X_{-1} can be defined equivalently as the dual of $(X^*)_1$, where $(X^*)_1 = D(A^*)$ with the graph norm. In this setting, X_1 and X_{-1} are investigated in Salamon [26] and also appear in Salamon [27]-[29], Lasiecka and Triggiani [17].

PROPOSITION 3.3. *With the notation of Definition* 3.1, *let* $\mu \in \rho(A)$ (*if* X *is real, take* $\mu \in \mathbb{R}$). *Then the operator*

$$R_\mu = (\mu I - A)^{-1}$$

has a (*unique*) *continuous extension to an operator in* $\mathcal{L}(X_{-1})$, *which we denote by the same symbol.* R_μ *is an isomorphism from* X_{-1} *to* X *and from* X *to* X_1.

If $L \in \mathcal{L}(X)$ *commutes with* A, *i.e., if*

$$LAx = ALx \quad \forall x \in D(A),$$

then the restriction of L *to* X_1 *belongs to* $\mathcal{L}(X_1)$ *and is the image of* L *via any of the isomorphisms* R_μ. *Further,* L *has a* (*unique*) *continuous extension to an operator in* $\mathcal{L}(X_{-1})$, *which is the image of* L *via any of the isomorphisms* R_μ^{-1}.

Proof. The fact that R_μ is an isomorphism from X_{-1} to X and from X to X_1 follows from Remark 3.2. The properties of L follow from the identities

$$Lx = R_\mu L R_\mu^{-1} x \quad \forall x \in D(A),$$

$$Lx = R_\mu^{-1} L R_\mu x \quad \forall x \in X. \qquad \square$$

Remark 3.4. Taking $L = \mathbb{T}_t$, $t \geqq 0$, we get from Proposition 3.3 that $\mathbb{T}$ has a restriction to a semigroup on X_1 whose generator is the restriction of A to $D(A^2)$ and $\mathbb{T}$ has an extension to a semigroup on X_{-1} whose generator is an extension of A, with domain X. Thus

$$A \in \mathcal{L}(X_1, X) \quad \text{and} \quad A \in \mathcal{L}(X, X_{-1}).$$

DEFINITION 3.5. Let X be a Banach space, let $\mathbb{T}$ be a semigroup on X with generator A, and let

$$f \in L^1_{\text{loc}}([0, \infty), X_{-1}).$$

Then we say that the function

$$x(\cdot) \in L^1_{\text{loc}}([0, \infty), X)$$

is a *strong solution* of the differential equation

$$\dot{x}(t) = Ax(t) + f(t) \tag{3.1}$$

if for any $t \geqq 0$

$$x(t) - x(0) = \int_0^t [Ax(s) + f(s)]\, ds.$$

If additionally, $x(\cdot)$ is continuous in X, i.e., if

$$x(\cdot) \in C([0, \infty), X),$$

then we say that $x(\cdot)$ is a *continuous state strong solution* of the differential equation above.

Our definition of strong solution follows that of Pazy [22, p. 109], if we replace X_{-1} by X and X by X_1. (His Definition 2.8 is slightly flawed; we must add an absolute continuity condition.) What we call a continuous state strong solution is called simply a solution by Salamon [26] (his Definition 3.3).

Remark 3.6. It is obvious that, as an X_{-1}-valued function, any strong solution of (3.1) is absolutely continuous and almost everywhere differentiable. However, since as an X-valued function it is only defined almost everywhere, it might happen that on a null set $x(t) \notin X$. For example, suppose X is a Hilbert space, $\mathbb{T}$ is analytic, and $f \in L^2_{\text{loc}}([0, \infty), X_{-1})$. Then for any initial condition $x(0) = x_0$, (3.1) has a strong solution $x(\cdot) \in L^2_{\text{loc}}([0, \infty), X)$ (see Lions and Magenes [20, Vol. II, p. 22] or Lasiecka [14, p. 325]). However, this strong solution might blow up in finite time with respect to the norm of X (see, for example, Lions [19, p. 202]).

Remark 3.7. In Definition 3.5, the condition

$$x(\cdot) \in L^1_{\text{loc}}([0, \infty), X)$$

is equivalent to

$$Ax(\cdot) \in L^1_{\text{loc}}([0, \infty), X_{-1})$$

and to

$$\dot{x}(\cdot) \in L^1_{\text{loc}}([0, \infty), X_{-1}).$$

If $0 \in \rho(A)$, then this is obvious, A being an isomorphism from X to X_{-1}. If A is not invertible, then to prove the equivalence we must take some $\lambda \in \rho(A)$, replace $x(t)$ by $y(t) = e^{-\lambda t} x(t)$, and do some simple computations.

Remark 3.8. In the conditions of Definition 3.5, if (3.1) has some strong solution, then for any $x_0 \in X$ it has a unique strong solution with $x(0) = x_0$ and that is given by the variation of parameters formula (see (1.2) with $f = Bu$).

To prove this statement it is most convenient to introduce the space X_{-2}, which is obtained from X_{-1} in the same way as X_{-1} is obtained from X. When we take X_{-1} as our new state space and X_{-2} as our new extended space, it is clear that *any strong solution in X is a continuous state strong solution in X_{-1}*. For continuous state strong

solutions, we can easily prove the statement by slightly adjusting a proof in Pazy [22, p. 105].

Now we can state the representation theorem.

THEOREM 3.9. *Let U and X be Banach spaces, let $p \in [1, \infty)$, and let $\Omega = L^p([0, \infty), U)$. Let $(\mathbb{T}, \Phi)$ be an abstract linear control system on X and Ω. Then there is a unique operator $B \in \mathcal{L}(U, X_{-1})$ such that for any $u \in \Omega$ and any $t \geqq 0$*

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma)\, d\sigma. \tag{3.2}$$

Moreover, for any $x_0 \in X$ and $u \in \Omega$ the function of $t \geqq 0$

$$x(t) = \mathbb{T}_t x_0 + \Phi_t u \tag{3.3}$$

is the (unique) continuous state strong solution of the differential equation

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{3.4}$$

with

$$x(0) = x_0.$$

Proof. For any $v \in U$ let us denote by ω_v the constant function on $[0, \infty)$ equal to v everywhere. Then ω_v belongs to the space $\tilde{\Omega}$ (see Remark 2.7) and so

$$\phi_v(t) = \Phi_t \omega_v$$

is a well-defined function from $[0, \infty)$ to X. Proposition 2.3 and causality (2.2) imply that $\phi_v(\cdot)$ is continuous. Using again causality (2.2) and the fact that $\|P_t \omega_v\| = t^{1/p}\|v\|$, we get from Proposition 2.5 that for suitable $\omega > 0$,

$$\|\phi_v(t)\| \leqq L\, e^{\omega t} t^{1/p} \|v\|.$$

It follows that for $s \in \mathbb{C}$ with $\operatorname{Re} s$ sufficiently big, the Laplace transform $\hat{\phi}_v(s)$ of ϕ_v is well defined.

The composition property (2.1), in the particular case of the input ω_v, means that for any $t, \tau \geqq 0$

$$\phi_v(t+\tau) = \mathbb{T}_t \phi_v(\tau) + \phi_v(t).$$

Applying the Laplace transformation with respect to t, we get

$$e^{s\tau}\hat{\phi}_v(s) - e^{s\tau}\int_0^\tau e^{-st}\phi_v(t)\, dt = (sI - A)^{-1}\phi_v(\tau) + \hat{\phi}_v(s),$$

or, rearranging and assuming $\tau > 0$, we have

$$\frac{e^{s\tau} - 1}{\tau}\hat{\phi}_v(s) = \frac{e^{s\tau}}{\tau}\int_0^\tau e^{-st}\phi_v(t)\, dt + (sI - A)^{-1}\frac{\phi_v(\tau)}{\tau}.$$

Taking the limit for $\tau \to 0$ (with respect to the norm of X) and using the continuity of ϕ_v and the fact that $\phi_v(0) = 0$ (Remark 2.2), we get

$$s\hat{\phi}_v(s) = \lim_{\tau \to 0} (sI - A)^{-1}\frac{\phi_v(\tau)}{\tau}, \tag{3.5}$$

in particular the limit on the right-hand side exists. Since $(sI - A)^{-1}$ is an isomorphism from X_{-1} to X (Proposition 3.3), it follows that with respect to the norm of X_{-1} the limit

$$Bv = \lim_{\tau \to 0} \frac{\phi_v(\tau)}{\tau} \tag{3.6}$$

exists and defines a linear operator from U to X_{-1}. We can rewrite (3.5) in the form

$$\hat{\phi}_v(s) = \frac{1}{s}(sI - A)^{-1}Bv, \tag{3.7}$$

which shows ($(sI-A)^{-1}$ being an isomorphism) that

$$B \in \mathscr{L}(U, X_{-1}).$$

The Laplace transformation being one-to-one on the space of continuous, exponentially bounded functions, (3.7) implies that

$$\phi_v(t) = \int_0^t \mathbb{T}_\sigma Bv \, d\sigma. \tag{3.8}$$

Next we show that if $u \in \Omega$ is a step function, then (3.2) holds. We proceed by induction after the number n of bounded intervals on which u is constant (which are of course followed by an unbounded interval on which $u(t) = 0$). For $n = 1$ we already know that the statement is true (see (3.8)). Suppose we know it is true for some $n \in \mathbb{N}$ and let $u \in \Omega$ be constant on $n+1$ bounded intervals, the last of these being $[\tau, T)$. Let $u_\tau(t) = u(\tau + t)$, so $u = u \underset{\tau}{\diamond} u_\tau$. We have for $t \geqq \tau$ (see (2.1))

$$\begin{aligned}
\Phi_t u &= \mathbb{T}_{t-\tau}\Phi_\tau u + \Phi_{t-\tau}u_\tau \\
&= \int_0^\tau \mathbb{T}_{t-\sigma}Bu(\sigma)\, d\sigma + \int_0^{t-\tau} \mathbb{T}_{t-\tau-\sigma}Bu(\tau+\sigma)\, d\sigma \\
&= \int_0^t \mathbb{T}_{t-\sigma}Bu(\sigma)\, d\sigma.
\end{aligned}$$

Thus we have proved (3.2) for any step function. The step functions being dense in Ω (recall that we have assumed $p < \infty$), it follows that (3.2) is true for any $u \in \Omega$.

The uniqueness of the operator B for which (3.2) holds is obvious.

Let us prove that $x(\cdot)$ given by (3.3) is the continuous state strong solution of (3.4) with $x(0) = x_0$. Using

$$\mathbb{T}_t x_0 - x_0 = \int_0^t A\mathbb{T}_s x_0 \, ds$$

the integral equation that $x(\cdot)$ must verify (see Definition 3.5) reduces to

$$\Phi_t u = \int_0^t A\Phi_s u \, ds + \int_0^t Bu(s)\, ds,$$

which is verified by an easy computation using the representation (3.2).

The fact that $x(\cdot) \in C([0, \infty), X)$ follows from Proposition 2.3. For uniqueness see Remark 3.8. □

PROBLEM 3.10. How much of Theorem 3.9 remains valid for $p = \infty$?

Remark 3.11. Salamon [28] has proved a representation theorem that partly overlaps our Theorem 3.9. It concerns systems that have input, state, and output (as opposed to our theorem, which considers only input and state). The part of Salamon's representation theorem that concerns the relationship between input and state is somewhat less general than our Theorem 3.9 in that X and U are assumed to be Hilbert spaces and $p = 2$. Further, the continuity of the state $x(t)$ as a function of the time t is assumed from the outset. Salamon's technique of proof is different from ours.

We mention another representation theorem which appears in Desch, Lasiecka, and Schappacher [7, p. 194]. There, it is proved that the solution of a well-posed initial boundary value problem is given by a formula equivalent to (1.2) of our paper.

Remark 3.12. It follows from Proposition 2.5 that for any $u \in \Omega$, $\Phi_t u$ has a Laplace transform $\hat{\Phi}_s u$. Using the representation (3.2) for $p < \infty$ we get that

$$\hat{\Phi}_s u = (sI - A)^{-1} B\hat{u}(s).$$

Remark 3.13. In the Introduction we have given a definition of admissible control operators. According to that definition, we can say that Theorem 3.9 yields a *necessary condition for admissibility* of B, namely $B \in \mathscr{L}(U, X_{-1})$. However, in the definition of admissibility, which we give at the beginning of § 4, we cut things short by demanding from the beginning $B \in \mathscr{L}(U, X_{-1})$, which makes the formulation easier.

Remark 3.14. If we assume from the outset that the system is modeled by (1.2), where $X \subset V$ with continuous and dense embedding, $B \in \mathscr{L}(U, V)$, $\mathbb{T}$ has a continuous extension to V and A has a continuous extension to an operator in $\mathscr{L}(X, V)$, then the necessary condition $B \in \mathscr{L}(U, X_{-1})$ can be obtained directly from the identity

$$B\mathrm{v} = \int_0^1 \mathbb{T}_t B\mathrm{v}\, dt - A \int_0^1 \mathbb{T}_{1-t} B\mathrm{v} t\, dt,$$

valid for any $\mathrm{v} \in U$ (proof by continuity, approximating $B\mathrm{v}$ by elements of $D(A)$). No reference to Theorem 3.9 is needed.

4. Spaces of admissible control operators. It is natural to ask whether the converse of Theorem 3.9 is true. More precisely, the problem is as follows. Let U and X be Banach spaces, $p \in [1, \infty)$ and $\Omega = L^p([0, \infty), U)$. Any abstract linear control system $(\mathbb{T}, \Phi)$ is, according to Theorem 3.9, completely determined by the pair (A, B), where A generates $\mathbb{T}$ and $B \in \mathscr{L}(U, X_{-1})$. However, as we shall see, not any such pair generates an abstract linear control system. This fact motivates the following definition.

DEFINITION 4.1. Let U and X be Banach spaces, let $p \in [1, \infty]$, and let $\Omega = L^p([0, \infty), U)$. Let $\mathbb{T}$ be a strongly continuous semigroup on X and let $B \in \mathscr{L}(U, X_{-1})$. For any $t \geqq 0$ we define the operator $\Phi_t : \Omega \to X_{-1}$ by

$$\Phi_t u = \int_0^t \mathbb{T}_{t-\sigma} B u(\sigma)\, d\sigma. \tag{4.1}$$

Then we say that B is *p-admissible* for $\mathbb{T}$ if for any $t \geqq 0$, $\Phi_t \in \mathscr{L}(\Omega, X)$.

It is easy to see that B is p-admissible if and only if the corresponding families of operators $(\mathbb{T}, \Phi)$ are an abstract linear control system on X and Ω (see Definition 2.1). For $p < \infty$ this is further equivalent with the fact that for any $u \in \Omega$ the differential equation (1.1) has a continuous state strong solution (see Definition 3.5). If the exponent p is clear from the context, we simply say that B is admissible.

Hypotheses equivalent to p-admissibility have appeared several times in the literature. For example, in Salamon [26] this condition is called hypothesis H2; in Salamon [29], S2; in Curtain and Salamon [5], H1; in Pritchard and Townley [24], A6. In Dolecki and Russell [9] and Ho and Russell [12] the condition appears in a dual form, essentially expressing that Φ_t^* is bounded. The following simple proposition permits us to relax the condition posed on B in Definition 4.1.

PROPOSITION 4.2. *Let U, X, p, Ω, and $\mathbb{T}$ be as in Definition* 4.1, *let $B \in \mathscr{L}(U, X_{-1})$, and let Φ_t be given by* (4.1). *If for some fixed $T > 0$ and any $u \in \Omega$*

$$\Phi_T u \in X,$$

then B is admissible.

Proof. With the notation of Proposition 3.3, let $B_0 = R_\mu B$. Then $B_0 \in \mathscr{L}(U, X)$ and we have

$$\Phi_T u = (\mu I - A) \int_0^T \mathbb{T}_{T-\sigma} B_0 u(\sigma)\, d\sigma,$$

which shows that Φ_T is closed. By the closed graph theorem Φ_T is bounded.

By (2.3) (see the proof of Proposition 2.3), Φ_t is bounded for all $t \leqq T$. The identity (2.4) (see the proof of Proposition 2.5), with rescaling, implies that if Φ_t is bounded for some t, it is bounded for all multiples of t. Therefore Φ_t is bounded for all $t \geqq 0$. □

For invertible semigroups, Proposition 4.2 admits the following strengthening, which is useful in applications.

PROPOSITION 4.3. *Let U, X, p, Ω, and $\mathbb{T}$ be as in Definition* 4.1, *let $B \in \mathscr{L}(U, X_{-1})$, and let Φ_t be given by* (4.1). *If $\mathbb{T}$ is invertible and if for some fixed $T > 0$ and any $u \in \Omega$*

$$\{t \geqq T \,|\, \Phi_t u \in X\} \neq \varnothing,$$

then B is admissible.

Proof. If B is not admissible, by Proposition 4.2, for any $T > 0$, we can find a $v \in \Omega$ such that $\Phi_T v \notin X$. Let $u \in \Omega$ be equal to v on $[0, T]$, and zero for $t > T$; then $\Phi_t u \notin X$ for any $t \geqq T$. □

Later we will prove a nontrivial strengthening of this proposition for the case $p < \infty$ (see Theorem 4.12). For semigroups that are not invertible, Proposition 4.3 is generally false (see Remark 3.6).

Remark 4.4. It happens that for certain A and B we can prove that for any $u \in \Omega$ (1.1) has a strong solution $x(\cdot)$ (see Definition 3.5), so $x(\cdot) \in L^1_{\text{loc}}([0, \infty), X)$, in particular $x(t)$ is almost everywhere in X. If $\mathbb{T}$ is invertible then Proposition 4.3 implies that B is admissible. In particular, if $p < \infty$ then, by Proposition 2.3, $x(\cdot)$ is continuous.

In other words, for $\mathbb{T}$ invertible, $p < \infty$, and $x(\cdot)$ given by (1.2), we have

$$x(\cdot) \in L^1_{\text{loc}}([0, \infty), X) \;\Rightarrow\; x(\cdot) \in C([0, \infty), X).$$

There are some other techniques to lift regularity from L^1_{loc} to C (see Lasiecka and Triggiani [16] and Lasiecka, Lions, and Triggiani [18]).

DEFINITION 4.5. Let U, X, p, and $\mathbb{T}$ be as in Definition 4.1. The space $\mathscr{B}_p(U, X, \mathbb{T})$ is the vector space of all p-admissible control operators B for the given U, X, and $\mathbb{T}$, with the norm

$$|||B|||_p = \sup_{\|u(\cdot)\|_{L^p([0,T],U)} \leqq 1} \left\| \int_0^T \mathbb{T}_{T-\sigma} B u(\sigma)\, d\sigma \right\|, \tag{4.2}$$

where $T > 0$ is fixed (see Remark 4.6 below).

We use the notation $|||\cdot|||$ (with three bars) to avoid confusion with the norm of B as an element of $\mathscr{L}(U, X_{-1})$. When U is just $\mathscr{K}$, the field of the scalars ($\mathbb{R}$ or $\mathbb{C}$), we denote

$$\ell_p(X, \mathbb{T}) = \mathscr{B}_p(\mathscr{K}, X, \mathbb{T}).$$

We usually write $\mathscr{B}_p$ and ℓ_p, without the arguments, when there is no danger of confusion. An operator from the scalars $\mathscr{K}$ to a space will always be identified with

an element of that space; in particular, elements of ℓ_p will be identified with elements of X_{-1}.

Remark 4.6. It is easy to verify, using (2.3) and (2.4), that for any different $T_1 > 0$ instead of T in (4.2) we get an equivalent norm and, using the representation theorem (Theorem 3.9), that $\mathcal{B}_p$ is complete for $p < \infty$ (for $p = \infty$, I do not know).

Remark 4.7. The following inclusions are immediate:

$$\mathcal{L}(U, X) \subset \mathcal{B}_p \subset \mathcal{L}(U, X_{-1}), \qquad X \subset \ell_p \subset X_{-1},$$

$$\mathcal{B}_{p_1} \subset \mathcal{B}_{p_2} \quad \text{for} \quad p_1 \leqq p_2, \tag{4.3}$$

all with continuous embedding. Further, it is clear that a *necessary condition* for $B \in \mathcal{B}_p$ is that the range of B satisfies

$$\operatorname{Ran} B \subset \ell_p \tag{4.4}$$

(this condition is not sufficient; see § 5).

It now seems plausible that $\mathcal{B}_p$ can be obtained as the completion of $\mathcal{L}(U, X)$ with respect to $|||\cdot|||_p$. But we shall see in § 5 that this is not the case.

There is one case when the determination of $\mathcal{B}_p$ is very easy, namely when $p = 1$ and X is reflexive.

THEOREM 4.8. *Let U, X, and $\mathbb{T}$ be as in Definition* 4.1 *and suppose X is reflexive. Then*

$$\mathcal{B}_1 = \mathcal{L}(U, X).$$

Proof. It will be enough to show that

$$\ell_1 = X. \tag{4.5}$$

Indeed, if (4.5) holds, then by (4.4) we have for any $B \in \mathcal{B}_1$ that $\operatorname{Ran} B \subset X$, and by the closed graph theorem, $B \in \mathcal{L}(U, X)$.

So let $b \in \ell_1$ and let the operator $\phi : L^1[0, 1] \to X_{-1}$ be given by

$$\phi v = \int_0^1 \mathbb{T}_\sigma b v(\sigma)\, d\sigma. \tag{4.6}$$

Because ϕ is obtained from Φ_1 by a change of variable in the integration, we have that, in fact, $\phi \in \mathcal{L}(L^1, X)$. X, being reflexive, has the Radon-Nikodym property (see Diestel and Uhl [8, pp. 76, 82]), so ϕ is representable (see [8, p. 63], i.e., there is some $g \in L^\infty([0, 1], X)$ such that for any $v \in L^1$

$$\phi v = \int_0^1 g(\sigma) v(\sigma)\, d\sigma.$$

When we compare the above equality with (4.6) it follows that $g(\sigma) = \mathbb{T}_\sigma b$ (a.e.), i.e., $\mathbb{T}_\sigma b \in L^\infty([0, 1], X)$ (but we do not know yet if $\mathbb{T}_\sigma b \in X$ for all σ, in particular, for $\sigma = 0$). From the identity

$$\frac{\mathbb{T}_t - I}{t} b = A \frac{1}{t} \int_0^t \mathbb{T}_\sigma b\, d\sigma$$

we get that for all $t \in (0, 1]$

$$\left\| \frac{\mathbb{T}_t - I}{t} b \right\|_{-1} \leqq \|A\| \cdot \|g\|$$

(by $\|A\|$ we mean the $\mathscr{L}(X, X_{-1})$ norm). By a theorem in Butzer and Berens [2, p. 88], b is in the domain of the generator of the semigroup $\mathbb{T}$ acting on X_{-1}, i.e., $b \in X$ (see Remark 3.4). □

Remark 4.9. In the above proof, once we have established that $\mathbb{T}_\sigma b$ is essentially bounded in X we could have alternatively used a lemma in Lions and Magenes [20, Vol. I, p. 275] to get the final result.

The above theorem is not true for general X, as we shall see in § 5.

Next we show that if a control operator is not admissible, then the state trajectory can be driven out of the state space using a smooth input on an arbitrarily short time-interval.

PROPOSITION 4.10. *Let U, X, p, and $\mathbb{T}$ be as in Definition* 4.1, *with $p<\infty$. Let $B \in \mathscr{L}(U, X_{-1})$ not be p-admissible. Then for any $T>0$ there is a function $u \in L^p[0, T]$ such that u is of class C^∞ on $[0, T)$ and*

$$\int_0^T \mathbb{T}_{T-\sigma} B u(\sigma)\, d\sigma \notin X. \tag{4.7}$$

Proof. Let $T>0$. Consider the Fréchet space

$$F = L^p[0, T] \cap C^\infty[0, T),$$

with the topology given by the increasing family of norms

$$q_n(u) = \|u\|_{L^p} + \sum_{k=0}^{n} \sup_{t \in [0,(1-(1/n))T]} \|u^{(k)}(t)\|,$$

where $n \in \mathbb{N}$. Let Φ_T denote the operator from $L^p[0, T]$ to X_{-1} defined by the left-hand side of (4.7). We have to show that $\Phi_T F$ is not contained in X. Suppose the contrary, then by the closed graph theorem $\Phi_T \in \mathscr{L}(F, X)$, i.e., there is some $n \in \mathbb{N}$ and some $K \geqq 0$ such that

$$\|\Phi_T u\| \leqq K \cdot q_n(u) \quad \forall u \in F.$$

In particular, for functions $u \in F$ the support of which is contained in $((1-(1/n))T, T)$, we have $q_n(u) = \|u\|_{L^p}$, so for such u

$$\left\| \int_{(1-(1/n))T}^{T} \mathbb{T}_{T-\sigma} B u(\sigma)\, d\sigma \right\| \leqq K \cdot \|u\|_{L^p}.$$

Making the change of variable $\sigma = (1-(1/n))T + s$, we get that for any $v \in C_0^\infty(0, T/n)$

$$\left\| \int_0^{T/n} \mathbb{T}_{(T/n)-s} B v(s)\, ds \right\| \leqq K \cdot \|v\|_{L^p}.$$

By the density of C_0^∞ in L^p (it is here that we need that $p<\infty$) we get that $B \in \mathscr{B}_p$, a contradiction. □

Now we give a necessary and sufficient condition for admissibility in the case when the semigroup is invertible. For that, we need the following remark concerning groups.

Remark 4.11. If the semigroup $\mathbb{T}$ acting on X is invertible, then for any (input) Banach space U and any $p \in [1, \infty]$, the admissible control operators for $\mathbb{T}$ and $\mathbb{T}^{-1}$ are the same, i.e.,

$$\mathscr{B}_p(U, X, \mathbb{T}) = \mathscr{B}_p(U, X, \mathbb{T}^{-1}).$$

The proof is very simple, by a change of variable.

THEOREM 4.12. *Let U, X, p, Ω, and $\mathbb{T}$ be as in Definition* 4.1, *with $p<\infty$, let $B\in\mathscr{L}(U, X_{-1})$ and let Φ_t be given by* (4.1). *If $\mathbb{T}$ is invertible and if for any $u\in\Omega$*

$$\{t>0 \mid \Phi_t u\in X\}\neq\varnothing, \tag{4.8}$$

then B is admissible.

Proof. Suppose B is not admissible. Then it is not admissible for $\mathbb{T}^{-1}$ either (see Remark 4.11), so by Proposition 4.10 there is some $v\in L^p[0,1]\cap C^\infty[0,1)$ for which $z\notin X$, where

$$z=\int_0^1 \mathbb{T}_{1-\sigma}^{-1}Bv(\sigma)\,d\sigma.$$

Let

$$u(s)=\begin{cases} v(1-s) & \text{for } s\in[0,1],\\ 0 & \text{for } s>1.\end{cases}$$

Then $\Phi_1 u=\mathbb{T}_1 z$, so $\Phi_1 u\notin X$. For any $t>0$ we have

$$\Phi_t u=\int_1^t \mathbb{T}_{t-\sigma}Bu(\sigma)\,d\sigma+\mathbb{T}_{t-1}\Phi_1 u.$$

The first term on the right-hand side is in X (by the smoothness of u), but the second is not. It follows that $\Phi_t u\notin X$, so condition (4.8) is not satisfied. □

I do not know if Theorem 4.12 remains true for $p=\infty$.

Remark 4.13. In the conditions of Theorem 4.12, if B is not admissible, then for any $T>0$ there is an L^p-function with support in $[0,T]$, of class C^∞ on $(0,\infty)$, such that $\Phi_t u\notin X$ for any $t>0$. To see that, we must multiply the function u constructed in the last proof by an appropriate cutoff function.

5. The wave equation on [0, π]. To deal with the controlled wave equation of Example 1.1, it will be helpful to analyse the periodic left-shift semigroup on $[0,2\pi]$ first. We introduce some notation.

We shall denote by $\mathbf{C}^\infty$ the Fréchet space of infinitely differentiable functions on the circle group $\mathbb{R}/2\pi\mathbb{Z}$ (periodic test functions). Following Edwards [10, Vol. II, p. 52], we shall denote by $\mathbf{D}$ the space of periodic distributions on $[0,2\pi]$, the dual of $\mathbf{C}^\infty$. For $s\in\mathbb{Z}$ and $r\in[1,\infty)$, $W_P^{s,r}$ will denote the periodic Sobolev space of order s and type r on $[0,2\pi]$, which we define as the space of those $\psi\in\mathbf{D}$ for which the Fourier transform $(\hat{\psi}_k)$ (which is a sequence over $\mathbb{Z}$) satisfies

$$(1+ik)^s\hat{\psi}_k=\hat{\varphi}_k,$$

where $i=\sqrt{-1}$ and $\varphi\in L^r[0,2\pi]$. For such ψ we set

$$\|\psi\|_{s,r}=\|\varphi\|_{L^r}.$$

For $r=2$ that becomes the more familiar

$$\|\psi\|_{s,2}=\sqrt{2\pi}\cdot\|(1+k^2)^{s/2}\hat{\psi}_k\|_{l^2}.$$

It is easy to prove, using propositions in Edwards [10, Vol. I, pp. 56, 114], that $W_P^{s,r}$ is decreasing with increasing s. It follows that for $s>0$, $W_P^{s,r}$ contains exactly those $\psi\in\mathbf{D}$ for which $\psi,\psi',\cdots,\psi^{(s)}$ are all in $L^r[0,2\pi]$.

Since the trigonometric polynomials are dense in $L^r[0,2\pi]$, it follows that they are dense in all the spaces $W_P^{s,r}$. In particular, for $s<0$, $W_P^{s,r}$ can be thought of as the completion of $W_P^{0,r}=L^r[0,2\pi]$ with respect to the norm $\|\cdot\|_{s,r}$.

For any combination of indices $1 \leqq p, r \leqq \infty$, (L^p, L^r) will denote the corresponding space of *Fourier multipliers*. The elements of (L^p, L^r) are functions $\beta : \mathbb{Z} \to \mathbb{C}$, for which the map $\psi \to \varphi$, $\hat{\varphi}_k = \beta_k \cdot \hat{\psi}_k$, is a bounded linear operator from $L^p[0, 2\pi]$ to $L^r[0, 2\pi]$. The norm of β is the norm of this operator. For more on multipliers see, for example, Edwards [10, Vol. II, Chap. 16].

Example 5.1. Let $1 \leqq r < \infty$. Let $\mathbf{S}$ be the semigroup of periodic left shifts on $Z = L^r[0, 2\pi]$, i.e.,

$$(\mathbf{S}_t z)(\zeta) = z(\zeta + t - k \cdot 2\pi) \quad \text{for} \quad k \cdot 2\pi \leqq \zeta + t < (k+1) \cdot 2\pi.$$

$\mathcal{A}$, the generator of $\mathbf{S}$, is given by

$$D(\mathcal{A}) = W_P^{1,r}, \qquad (\mathcal{A}z)(\zeta) = z'(\zeta).$$

The eigenvalues of $\mathcal{A}$ are ik, $k \in \mathbb{Z}$, and the corresponding eigenfunctions are $e^{ik\zeta}$, so $\mathbf{S}$ can be expressed in terms of Fourier series by

$$\widehat{(\mathbf{S}_t z)}_k = e^{ikt} \hat{z}_k.$$

It is easy to check that for this semigroup $Z_{-1} = W_P^{-1,r}$.

We want to determine the space $\mathcal{B}_p(Z, \mathbf{S})$ of admissible control operators for scalar L^p-inputs.

PROPOSITION 5.2. *For the semigroup above and for any index p with $1 \leqq p \leqq \infty$, we have*

$$\mathcal{B}_p(Z, \mathbf{S}) = \{b \in \mathbf{D} \mid \hat{b} \in (L^p, L^r)\},$$
$$|||b|||_p = 2\pi \cdot \|\hat{b}\|_{(L^p, L^r)}.$$

Proof. First we prove that

$$\{b \in \mathbf{D} \mid \hat{b} \in (L^p, L^r)\} \subset Z_{-1}.$$

Let b be in the set on the left-hand side above. Then obviously $\hat{b} \in l^\infty$. It follows that for any $\nu \in (1, 2]$, $(1+ik)^{-1}\hat{b}_k \in l^\nu$. By the Hausdorff-Young Theorem (Edwards [10, Vol. II, p. 153]) we get that $(1+ik)^{-1}\hat{b}_k = \hat{\varphi}_k$, where $\varphi \in L^\mu[0, 2\pi]$ for any $\mu \in [2, \infty)$, and hence for any $\mu \in [1, \infty)$. In particular, $\varphi \in L^r[0, 2\pi] = Z$, so $b \in W_P^{-1,r} = Z_{-1}$.

Next we take $b \in Z_{-1}$. It is easy to check that for any $u \in L^p[0, 2\pi]$,

$$z = \int_0^{2\pi} \mathbf{S}_{2\pi-\sigma} b u(\sigma)\, d\sigma \quad \text{iff} \quad \hat{z}_k = 2\pi \hat{b}_k \cdot \hat{u}_k. \tag{5.1}$$

Using Proposition 4.2 with $T = 2\pi$, we bet that $b \in \mathcal{B}_p(Z, \mathbf{S})$ if and only if $\hat{b} \in (L^p, L^r)$. Moreover, choosing $T = 2\pi$ in the definition of $|||\cdot|||_p$,

$$|||b|||_p = \sup_{\|u\|_{L^p}=1} \|z\|_{L^p} = 2\pi \|\hat{b}\|_{(L^p, L^r)}. \qquad \square$$

Generally, it is not easy to describe the spaces (L^p, L^r), but there are cases when it is. One such case is $r \leqq 2 \leqq p$. Then $(L^p, L^r) = l^\infty$ (see Edwards [10, Vol. II, p. 301]). Periodic distributions b for which $\hat{b} \in l^\infty$ are called pseudomeasures and following Edwards [10, Vol. II, p. 108], we shall denote their space by P. The norm on P is $\|b\|_P = \|\hat{b}\|_{l^\infty}$. So we have for $r \leqq 2 \leqq p$,

$$\mathcal{B}_p(Z, \mathbf{S}) = P,$$

with equivalent norms.

The completion of Z with respect to $|||\cdot|||_p$, still considering $r \leqq 2 \leqq p$, is the space of periodic distributions b for which $\hat{b} \in c_0$, where c_0 is the space of sequences convergent to zero for $|k| \to \infty$. But this is only a closed subspace of P. Therefore we have the following.

NEGATIVE RESULT 5.3. The state space Z may be a nondense subspace of $\ell_p(Z, \mathbf{S})$. In other words, in the space $\ell_p(Z, \mathbf{S})$, admissible unbounded control operators cannot always be approximated by bounded ones.

NEGATIVE RESULT 5.4. If Z is not reflexive, $\ell_1(Z, \mathbf{S})$ may contain unbounded elements (i.e., which are not in Z) (cf. Theorem 4.8).

That can be seen taking $p = r = 1$ in Proposition 5.2. Indeed, (L^1, L^1) consists of the Fourier transforms of elements of M, the space of regular bounded Borel measures ("Radon measures") on $[0, 2\pi)$ (see Edwards [10, Vol. II, pp. 53, 289]). Hence $\ell_1(Z, \mathbf{S}) = M$, which contains $L^1[0, 2\pi]$ strictly.

We give a negative result concerning reachability.

NEGATIVE RESULT 5.5. It is possible to have elements $z \in Z$ that are *completely unreachable*. By that we mean

$$z \neq \int_0^t \mathbf{S}_{t-\sigma} b u(\sigma)\, d\sigma$$

for any $t > 0$, any $b \in \ell_p(Z, \mathbf{S})$ and any $u \in L^p[0, t]$.

Indeed, consider the case $r \leqq 2 \leqq p$ in Example 5.1. Then, as we have seen, $\ell_p(Z, \mathbf{S}) = P$, which is independent of r and p. Therefore the reachable states are contained in the smallest of the state spaces for the various $r \leqq 2$, namely in $L^2[0, 2\pi]$. If $r < 2$, the state space $Z = L^r[0, 2\pi]$ is larger than that and therefore contains completely unreachable elements.

We mention that for $r = 2$, all elements of Z are reachable, for any $p \in [2, \infty)$. Indeed, for any given $z \in L^2[0, 2\pi]$ and $p \in [2, \infty)$, we can choose $b \in P$ with $\hat{b}_k = \pm 1$ and $u \in L^p[0, 2\pi]$ such that $\hat{z}_k = 2\pi \hat{b}_k \cdot \hat{u}_k$. That follows from a theorem in Edwards [10, Vol. II, p. 220], about changing signs of Fourier coefficients. According to (5.1), z is reachable.

The equality

$$\mathscr{B}_p(U, Z, \mathbf{S}) = \mathscr{L}(U, \ell_p(Z, \mathbf{S})) \tag{5.2}$$

is true for finite-dimensional U, and simple reasoning shows that if the equality $\mathscr{B}_p(U, Z, \mathbf{S}) = \mathscr{L}(U, V)$ holds for some $p \in [1, \infty]$ and some Banach spaces U and V, then $V = \ell_p(Z, \mathbf{S})$. However, we have the following negative result, already mentioned in the Introduction.

NEGATIVE RESULT 5.6. The representation (5.2) does not hold in general.

Indeed, take $r = p = 2$, $U = \ell_2(Z, \mathbf{S}) = \{b \in \mathbf{D} \mid \hat{b} \in l^\infty\}$, and $B = I$ (the identity on U). Let $u_0 \in U \backslash Z$ with $\hat{u}_0 \in c_0$. Then $u(t) = \mathbf{S}_t u_0$ is a continuous U-valued function and

$$\int_0^{2\pi} \mathbf{S}_{2\pi-\sigma} B u(\sigma)\, d\sigma = 2\pi \cdot u_0,$$

so B is not admissible.

Now we turn to the controlled one-dimensional wave equation. The notation X, A, $\mathbb{T}$, r, p, $W_0^{1,r}$, $W^{2,r}$ will have the same meaning as in Example 1.1. In addition, the notation of Example 5.1 still holds. We want to determine the space $\ell_p(X, \mathbb{T})$ of admissible one-dimensional control operators.

Using an isomorphism, we shall reduce this problem to that discussed in Example 5.1, with the minor difference that the coefficient with index $k = 0$ in all Fourier transforms must be set to zero. Let

$$\mathbf{D}^0 = \{\psi \in \mathbf{D} \mid \hat{\psi}_0 = 0\}.$$

The spaces

$$Z^0 = Z \cap \mathbf{D}^0, \qquad Z^0_{-1} = Z_{-1} \cap \mathbf{D}^0$$

are closed subspaces of Z and Z_{-1}, respectively. Let the isomorphism $H: Z^0 \to X$ be given by

$$(Hz)(\zeta) = \begin{pmatrix} \int_0^\zeta [z(\xi) + z(2\pi - \xi)]\, d\xi \\ z(\zeta) - z(2\pi - \zeta) \end{pmatrix} \quad \forall \zeta \in [0, \pi].$$

The operator H^{-1} is given by

$$(H^{-1}x)(\zeta) = \begin{cases} \frac{1}{2}[x_1'(\zeta) + x_2(\zeta)] & \text{for} \quad \zeta \in [0, \pi], \\ \frac{1}{2}[x_1'(2\pi - \zeta) - x_2(2\pi - \zeta)] & \text{for} \quad \zeta \in (\pi, 2\pi]. \end{cases}$$

The image of $\mathbb{T}$ through H^{-1} is the semigroup $\mathbf{S}^0$ on Z^0 generated by

$$\mathscr{A}^0 = H^{-1}AH = d/d\zeta,$$
$$D(\mathscr{A}^0) = H^{-1}D(A) = D(\mathscr{A}) \cap \mathbf{D}^0.$$

Hence $\mathbf{S}^0$ is the restriction of $\mathbf{S}$ to Z^0, i.e., the periodic left-shift semigroup on Z^0. Intuitively, the restriction of z to $[0, \pi]$ represents the left-bound component of x_2 and the restriction of z to $[\pi, 2\pi]$ represents the right-bound component of x_2, with opposite sign and reversed. The eigenvectors of A are

$$H e^{ik\zeta} = 2i \begin{pmatrix} \frac{1}{ik} \sin k\zeta \\ \sin k\zeta \end{pmatrix} \qquad \text{for} \quad k \in \mathbb{Z} \backslash \{0\}.$$

Since $(sI - \mathscr{A}^0)^{-1} = H^{-1}(sI - A)^{-1}H$, the operator H can be extended (uniquely) to an isomorphism from Z^0_{-1} to X_{-1}, still denoted H.

We get the general characterization

$$\ell_p(X, \mathbb{T}) = H\ell_p(Z^0, \mathbf{S}^0) = H\{\beta \in \mathbf{D}^0 \,|\, \hat{\beta} \in (L^p, L^r)\}. \tag{5.3}$$

For example, for $r \leqq 2 \leqq p$,

$$\ell_p(X, \mathbb{T}) = H\{\beta \in \mathbf{D}^0 \,|\, \hat{\beta} \in l^\infty\}.$$

We give a more concrete description of the operator H appearing in (5.3), defined only as a (hard to visualize) continuous extension. For this we need some notation.

For periodic test functions $\varphi \in \mathbf{C}^\infty$ we shall denote $\check{\varphi}(\zeta) = \varphi(2\pi - \zeta)$. For $\psi \in \mathbf{D}$, $\check{\psi}$ denotes the distribution defined by $\langle \varphi, \check{\psi} \rangle = \langle \check{\varphi}, \psi \rangle$. A distribution $\psi \in \mathbf{D}$ is called *antisymmetric* if $\check{\psi} = -\psi$. The antisymmetric elements of $\mathbf{D}$ form a closed subspace of $\mathbf{D}$, which we denote $\mathbf{D}^A$.

We introduce the test function space

$$\mathscr{D} = \left\{ \varphi \in C^\infty[0, \pi] \,\middle|\, \frac{d^k}{d\zeta^k} \varphi(0) = \frac{d^k}{d\zeta^k} \varphi(\pi) = 0, \quad \text{for } k = 0, 2, 4, \cdots \right\},$$

with the topology of uniform convergence of all derivatives. Any $\varphi \in \mathscr{D}$ has a (unique) *antisymmetric extension to* $[0, 2\pi]$, denoted Aext φ, which is an element of $\mathbf{C}^\infty$. The operator Aext is an isomorphism from $\mathscr{D}$ onto the antisymmetric part of $\mathbf{C}^\infty$.

Any $\psi \in \mathbf{D}^A$ has a natural *restriction to* $[0, \pi]$, denoted Rest ψ, which is an element of $\mathscr{D}'$, the dual of $\mathscr{D}$. Rest ψ is defined by

$$\langle \varphi, \text{Rest } \psi \rangle = \tfrac{1}{2} \langle \text{Aext } \varphi, \psi \rangle.$$

If the distribution ψ is represented by an integrable function, then Rest ψ, as defined above, coincides with the usual restriction. The operator Rest is an isomorphism from $\mathbf{D}^A$ onto $\mathscr{D}'$.

On $\mathbf{D}^0$, as opposed to $\mathbf{D}$, we can integrate distributions freely. That means that for any $\psi \in \mathbf{D}^0$ there is a unique $\psi_1 \in \mathbf{D}^0$ such that $\psi_1' = \psi$. We shall write $\psi_1 = \int \psi$.

We define the isomorphism $H_1 : \mathbf{D}^0 \to \mathbf{D}^A \times \mathbf{D}^A$ by

$$H_1\psi = \begin{pmatrix} \int \psi + \int \check{\psi} \\ \psi - \check{\psi} \end{pmatrix}.$$

It is easy to check that for any $z \in Z^0$

$$Hz = \text{Rest } H_1 z.$$

Since Rest H_1 is an isomorphism from $\mathbf{D}^0$ onto $\mathscr{D}' \times \mathscr{D}'$, the formula above must also hold for elements z in the extended space Z^0_{-1}.

We give now a more concrete description of X_{-1}.

On $\mathscr{D}$, the operator $d/d\zeta$ makes no sense, but $d^2/d\zeta^2$ is well defined and even invertible. By duality, this operator can be defined on $\mathscr{D}'$:

$$\left\langle \varphi, \frac{d^2}{d\zeta^2}\psi \right\rangle = \left\langle \frac{d^2}{d\zeta^2}\varphi, \psi \right\rangle.$$

If ψ happens to be twice differentiable and $\psi(0) = \psi(\pi) = 0$, then the definition above coincides with that of the usual second derivative.

If we denote by $W^{-1,r}[0, \pi]$ the image of $W_0^{1,r}[0, \pi]$ through $d^2/d\zeta^2$, then

$$X_{-1} = \bigtimes_{W^{-1,r}}^{L^r}.$$

Acknowledgments. I am grateful to Professor Zvi Artstein for his generous help and advice during the preparation of this paper. Professors Ruth Curtain, Irena Lasiecka, Anthony Pritchard, and Dietmar Salamon, through their letters, encouraged me and drew my attention to a number of papers of which I was not aware. The referees corrected some errors, and one pointed out reference [28] and suggested Remark 3.14.

REFERENCES

[1] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*, Grundlehren Math. Wiss., 223, Springer-Verlag, Berlin, New York, 1976.

[2] P. L. Butzer and H. Berens, *Semi-Groups of Operators and Approximation*, Grundlehren Math. Wiss. Einzeldarstell., 145, Springer-Verlag, Berlin, New York, 1967.

[3] R. F. Curtain, *On semigroup formulations of unbounded observations and control action for distributed systems*, in Proc. Mathematical Theory of Networks and Systems Symposium, Beer-Sheva, Israel, 1983, Springer-Verlag, Berlin, New York, 1984.

[4] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Information Sciences, 8, Springer-Verlag, Berlin, New York, 1978.

[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] G. Da Prato, *Abstract differential equations and extrapolation spaces*, in Proc. Conference on Infinite-Dimensional Systems, Retzhof, Federal Republic of Germany, 1983, Lecture Notes in Mathematics, 1076, Springer-Verlag, Berlin, New York, 1984, pp. 53–61.

[7] W. Desch, I. Lasiecka, and W. Schappacher, *Feedback boundary control problems for linear semigroups*, Israel J. Math., 51 (1985), pp. 177–207.

[8] J. Diestel and J. J. Uhl, *Vector Measures*, American Mathematical Society, Providence, RI, 1977.

[9] S. Dolecki and D. L. Russell, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185–220.

[10] R. E. EDWARDS, *Fourier Series, a Modern Introduction*, Vols. I and II, Graduate Texts in Mathematics, Springer-Verlag, Berlin, New York, 1982.

[11] H. O. FATTORINI, *Boundary control systems*, SIAM J. Control, 6 (1968), pp. 349-385.

[12] L. F. HO AND D. L. RUSSELL, *Admissible input elements for systems in Hilbert space and a Carleson measure criterion*, SIAM J. Control Optim., 21 (1983), pp. 614-640.

[13] R. E. KALMAN, P. L. FALB, AND M. A. ARBIB, *Topics in Mathematical Systems Theory*, McGraw-Hill, New York, 1969.

[14] I. LASIECKA, *Unified theory for abstract parabolic boundary problems—A semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287-333.

[15] I. LASIECKA AND R. TRIGGIANI, *A cosine operator approach to modeling $L_2(0, T; L_2(\Gamma))$-boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-93.

[16] ———, *Regularity of hyperbolic equations under $L_2(0, T; L_2(\Gamma))$-Dirichlet boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[17] ———, *Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations*, J. Differential Equations, 47 (1983), pp. 246-272.

[18] 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.

[19] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Grundlehren Math. Wiss. Einzeldarstell., 170, Springer-Verlag, Berlin, New York, 1971.

[20a] J. L. LIONS AND E. MAGENES, *Non Homogeneous Boundary Value Problems and Applications*, Vol. I, Grundlehren Math. Wiss. Einzeldarstell., 181, Springer-Verlag, Berlin, New York, 1972.

[20b] ———, *Non Homogeneous Boundary Value Problems and Applications*, Vol. II, Grundlehren Math. Wiss. Einzeldarstell., 182, Springer-Verlag, Berlin, New York, 1972.

[21] R. NAGEL, ED., *One-parameter Semigroups of Positive Operators*, Lecture Notes in Mathematics, 1184, Springer-Verlag, Berlin, New York, 1986.

[22] A. PAZY, *Semigroups of Linear Operators and Applications to PDE's*, Appl. Math. Sci., 44, Springer-Verlag, Berlin, New York, 1983.

[23] A. J. PRITCHARD AND A. WIRTH, *Unbounded control and observation systems and their duality*, SIAM J. Control Optim., 16 (1978), pp. 535-545.

[24] A. J. PRITCHARD AND S. TOWNLEY, *A stability radius for infinite dimensional systems*, in Proceedings, Vorau, Austria, 1986, Lecture Notes in Control and Information Sciences 102, Springer-Verlag, Berlin, New York, 1987.

[25] D. L. RUSSELL, *A Unified Boundary Controllability Theory for Hyperbolic and Parabolic Partial Differential Equations*, Stud. Appl. Math., 52, 1973, pp. 189-211.

[26] D. SALAMON, *Control and Observation of Neutral Systems*, Research Notes in Math., 91, Pitman, Boston, London, 1984.

[27] D. SALAMON, *An abstract framework for infinite dimensional systems with unbounded control and observation*, in Proceedings, Vorau, Austria, 1984, Lecture Notes in Control and Information Sciences 75, Springer-Verlag, Berlin, New York, 1985.

[28] ———, *Realization theory in Hilbert space*, Tech. Summary Report 2835, University of Wisconsin, Madison, WI, 1985. (Revised version submitted in 1988.)

[29] ———, *Infinite Dimensional Systems with Unbounded Control and Observation: A Functional Analytic Approach*, Trans. Amer. Math. Soc., 300 (1987), pp. 383-431.

[30] D. WASHBURN, *A bound on the boundary input map for parabolic equations with applications to time optimal control*, SIAM J. Control Optim., 17 (1979), pp. 652-671.

[31] G. WEISS, *Admissibility of input elements for diagonal semigroups on l^2*, Systems and Control Lett., 10 (1988), pp. 79-82.

[32] G. WEISS, *Admissible observation operators for linear semigroups*, Israel J. Math., 64 (1989), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 546-562, May 1989

006

APPROXIMATIONS AND OPTIMAL CONTROL FOR THE PATHWISE AVERAGE COST PER UNIT TIME AND DISCOUNTED PROBLEMS FOR WIDEBAND NOISE-DRIVEN SYSTEMS*

HAROLD J. KUSHNER†

Abstract. The average cost per unit time problem for wide bandwidth noise-driven control systems is considered, where the average cost is in the pathwise sense; no expectations are used. Let t = time of control and BW = bandwidth. For the class of processes considered here, various unformity properties are proven for the convergence of the pathwise average costs as $t \to \infty$, BW $\to \infty$. Let $u^\delta(\cdot)$ be a smooth δ-optimal control for the limit controlled diffusion (the limit as BW $\to \infty$) for the (mean) average cost per unit time problem. It is shown that for large enough t and BW, $u^\delta(\cdot)$ is 2δ-optimal (with a probability arbitrarily close to one) for the pathwise wide bandwidth problem. This uniformity is important in applications, for often there is only one long sequence to control, and the expectation is inappropriate. Also, otherwise, as BW $\to \infty$, it might take longer and longer to approximate the limit pathwise average cost well. Applications to related "pathwise average" problems are given: the convergence of the average pathwise errors for an "approximate" nonlinear filter with wide bandwidth observation and system driving noise, and the convergence and accuracy of Monte Carlo calculations of Lyapunov exponents for wide bandwidth noise-driven systems (as BW $\to \infty$) via average cost/unit time methods. It is also shown for the discounted cost problem that the optimum pathwise costs converge to the minimum average cost per unit time as both the discount factor goes to zero and BW $\to \infty$.

Key words. pathwise average cost per unit time, ergodic control, approximations of ergodic control, wide band noise driven systems, approximate nonlinear filtering, Lyapunov exponents, discounted cost

AMS(MOS) subject classifications. 93E20, 93E15, 93E11, 60F17

1. Introduction.

Average cost per unit time (over an infinite time horizon) optimal control problems for diffusion and other Markov models have been dealt with in various ways, as in, e.g., [1]-[3]. We treat such a problem for "wideband, noise-driven," and related systems, which are "close" to a diffusion, and when the average is in the pathwise but not necessarily in the mean value sense. The general method works for many other classes of processes that are suitably approximated by an appropriate controlled Markov process. As is pointed out below and in §§ 4 and 5, the results have applications to many other problems where pathwise averages are important, and the noises are "wide band." For example, in § 5 we treat the problem where both BW $\to \infty$ and the discount factor goes to zero.

Let the diffusion model be given in the relaxed control form:

$$dx = \left[\int \bar{b}(x, \alpha) m_t(d\alpha)\right] dt + \sigma(x)\, dw, \qquad x \in R^r, \quad \text{Euclidean } r\text{-space} \tag{1.1}$$

where $\bar{b}(\cdot, \cdot)$ and $\sigma(\cdot)$ are continuous (other conditions will be listed below) and $m_t(\cdot)$ is an admissible relaxed control [1], [3], [4], over a compact control value space U. The relaxed control might be of the feedback form. The $w(\cdot)$ is a standard vector-valued Wiener process, and the dimensions of the vectors b, w, and σ are compatible.

* Received by the editors January 27, 1988; accepted for publication (in revised form) July 6, 1988. This work was supported in part by Air Force Office of Scientific Research grant 85-0315, Army Research Office grant DAAL03-86-K-0171, and Office of Naval Research grant N00014-85-K-0607.

† Lefschetz Center for Dynamical Systems, Department of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

DEFINITION. Relaxed Control. Let U be a compact set in some Euclidean space. Let the $w(\cdot)$ in (1.1) be a Wiener process with respect to a filtration $\{\mathcal{F}_t\}$. A measure valued (a measure on the Borel sets of $U\times[0,\infty)$) random variable $m(\cdot)$ is an admissible relaxed control if $\iint_0^t f(s,\alpha)m(ds\,d\alpha)$ is progressively measurable for each bounded and continuous $f(\cdot)$ and $m([0,t]\times U)=t$, for all $t\geqq 0$. If $m(\cdot)$ is admissible, then there is a derivative $m_t(\cdot)$ (defined for almost all t) that is nonanticipative and

$$\int_0^t\int f(s,\alpha)m(ds\,d\alpha)=\int_0^t ds\int f(s,\alpha)m_s(d\alpha)$$

for all t with probability one (w.p.1). Sometimes we use the "feedback" relaxed control (which we write as $m_x(\cdot)$), which is a measure on the Borel sets of U for each x, and $m_x(B)$ is Borel-measurable for each Borel B. The $m_t(\cdot)$ and $m_x(\cdot)$ will also be referred to as relaxed controls.

In [1], relaxed controls have been used to get nearly optimal controls for several "wideband" noise-driven systems, and in [3], they have been used cleverly to get an "occupation measure" for the state-control pair, which ultimately allowed the authors to demonstrate the existence of an optimal stationary control. These advantages also occur for the particular problems described below. In [1] and [2], the cost of concern is ([2] does not use relaxed controls)

$$\bar{\gamma}(m)=\overline{\lim_T}\frac{1}{T}\int_0^T\int Ek(x(t),\alpha)m_t(d\alpha)\,dt \tag{1.2}$$

for a bounded continuous $k(\cdot)$.

In practice, of course, we do not have a process that is a diffusion, and it is of considerable interest to consider systems of the following form:

$$\dot{x}^\varepsilon=\int b(x^\varepsilon,\alpha)m_t(d\alpha)+F_\varepsilon(x^\varepsilon,\xi^\varepsilon), \tag{1.3}$$

where $\xi^\varepsilon(\cdot)$ is a wide bandwidth noise process and we wish to minimize the following:

$$\bar{\gamma}^\varepsilon(m)=\overline{\lim}\frac{1}{T}\int_0^T\int Ek(x^\varepsilon(t),\alpha)m_t(d\alpha)\,dt. \tag{1.4}$$

For convenience in the development (to simplify the details), we use a process defined by the scaling $\xi^\varepsilon(t)=\xi(t/\varepsilon^2)$, where the "primitive" process $\xi(\cdot)$ satisfies certain mixing conditions (one of the four sets (A2.3), (A2.4), (A2.5) or (A2.10) or (A3.2), (A3.3), (A3.4) or (A3.5) below). This scaling is, in fact, a common way of constructing a wide bandwidth process, and is also used in [13] and [14]. But it should be clear from the development that the method is much more generally applicable. Reference [1] has dealt with a system of type (1.3) (with weak limit of type (1.1)) and cost of the form (1.2). It has been shown, under the conditions there, that for any $\delta>0$, a smooth δ-optimal control u^δ for (1.1), (1.2) was also "nearly" optimal for (1.3) and (1.4), for small ε in the sense that $\underline{\lim}_\varepsilon \bar{\gamma}^\varepsilon(m^\varepsilon)\geqq\lim_\varepsilon\bar{\gamma}^\varepsilon(u^\delta)-\delta$ for any sequence m^ε. Such results are helpful in justifying the use of the ideal limit process (1.1) for use in control theory.

In [3], Borkar and Ghosh have shown the existence of an optimal feedback control for the diffusion model (under this control the diffusion could be taken to be stationary) and cost function (1.2), but *with the E deleted*—a pathwise result. This paper is devoted to a related problem for the model (1.3). Define

$$\gamma_T(m)=\frac{1}{T}\int_0^T\int k(x(s),\alpha)m_s(d\alpha)\,ds,\qquad \gamma(m)=\overline{\lim_T}\,\gamma_T(m), \tag{1.5}$$

$$\gamma_T^\varepsilon(m) = \frac{1}{T}\int_0^T \int k(x^\varepsilon(s), \alpha) m_s(d\alpha)\, ds. \tag{1.6}$$

If $m(\cdot)$ is equivalent to a classical control function $u(\cdot)$, we write u in lieu of m in $\gamma_T(m)$, etc. The "pathwise" convergence result in [3] is of particular importance in applications, since we often have a single long realization, and then the expectation is not appropriate in the cost function. The results in [3] (under their conditions) give the existence of a feedback relaxed control $\bar{m}(\cdot)$ such that

$$\gamma_T(\bar{m}) \to \gamma = \inf_m \cdot \overline{\lim_T}\, \gamma_T(m) \quad \text{w.p.1.} \tag{1.7}$$

In our problem here, owing to the wideband noise and the appearance of the two parameters ε and T, convergence results of the "almost sure" type are often rather meaningless from a practical point of view as well as being nearly impossible to obtain. They might have little meaning for the following reason. Typically, in an application we have a particular process with a given wide bandwidth driving force. We are interested in knowing how well good controls for the "limit" problem do on the actual "physical" problem as well as various qualitative properties of the "physical" process. The wide bandwidth driving term is *imbedded into a sequence* for the purpose of getting such an approximation result, and "almost sure"-type results might have little practical importance.

Let $u^\delta(\cdot)$ denote a "nice" δ-optimal classical control ("nice" is defined in the next section) for model (1.1) and cost function (1.4). Then we wish to show:

$$\gamma_T^\varepsilon(u^\delta) \xrightarrow{P} \bar{\gamma}(u^\delta) \quad \text{as } \varepsilon \to 0, \quad T \to \infty, \tag{1.8a}$$

$$\underline{\lim_{\varepsilon, T}}\, P\{\gamma_T^\varepsilon(m^\varepsilon) \geqq \bar{\gamma}(u^\delta) - \delta\} = 1 \tag{1.8b}$$

for any sequence of admissible relaxed controls $m^\varepsilon(\cdot)$. Since the time derivative of $\gamma_T^\varepsilon(m)$ is $O(1/T)$ uniformly in ε, m, ω, the convergence is somewhat stronger than indicated by (1.8). Equation (1.8b) implies a type of uniformity of convergence, since the way that $\varepsilon \to 0$ and $T \to \infty$ is not important. Were this "uniformity" not the case, it would be possible that as $\varepsilon \to 0$, a larger and larger T is needed to closely approximate the limit value. In that case, the white noise limit (1.1) would not be useful for predictive or control purposes when the true model is (1.3).

In § 2, we list several assumptions and prove (1.8). To simplify the development, the technique of perturbed test functions from [5] is used. To facilitate the calculations, some of the conditions will be adapted from those used in that reference—but many useful generalizations should be clear. In § 3, we redevelop the result of § 2, using a "first-order perturbed test function" method, with less smoothness required on the functions and less mixing required on the noise but more details required in the proof. Some extensions are discussed in § 4. The ideas of "pathwise uniform" convergence of a sample average cost per unit time have many other applications; for example in the Monte Carlo evaluation of Lyapunov exponents with wide bandwidth noise coefficients for linear systems [6]. The formula for the Lyapunov exponent is of the form of an average cost per unit time. For this problem, it is shown in § 4 that the Monte Carlo-evaluated pathwise average cost per unit time converges (as $\varepsilon \to 0$, $T \to \infty$) to the same limit that we would obtain were the actual limit diffusion used for the evaluation. The limit depends only on the correlation function of the noise $\xi^\varepsilon(\cdot)$. Such a result is essential for the Monte Carlo method to be useful and for the Lyapunov exponents of the limit system to be meaningful indicators of the behavior of the actual (wide bandwidth, noise-driven) physical system.

An extension to a problem of average pathwise error per unit time for an "approximate" nonlinear filter for a system with wide bandwidth driving and observation noise is also discussed in § 4.

In § 5, we treat extensions to the discounted cost case. Define the pathwise discounted cost

$$V_\beta^\varepsilon(m) = \beta \int_0^\infty e^{-\beta s} \int k(x^\varepsilon(s), \alpha) m_s(d\alpha)\, ds,$$

and let $m^\varepsilon(\cdot)$ be a sequence of "δ_1-optimal" controls. We show that, for $u^\delta(\cdot)$ defined as above (1.8):

$$V_\beta^\varepsilon(u^\delta) \xrightarrow{P} \bar{\gamma}(u^\delta) \quad \text{as } \beta \to 0, \quad \varepsilon \to 0, \tag{1.9a}$$

$$\varliminf_{\varepsilon,\beta} P\{V_\beta^\varepsilon(m^\varepsilon) \geqq \bar{\gamma}(u^\delta) - \delta\} = 1. \tag{1.9b}$$

The uniformity result is important, since we would not want the speed with which $\beta \to 0$ to depend on the bandwidth—to get the proper approximation. The sense in which $m^\varepsilon(\cdot)$ is δ_1-optimal is purposely left vague—since (1.9) holds for any $\{m^\varepsilon(\cdot)\}$, under the conditions below. Thus for small ε, β, $u^\delta(\cdot)$ is always nearly optimal. There also are extensions to impulsive and singular control problems.

2. A basic convergence theorem. For convenience in this section, we use the assumptions of [5, Chap. 4.6], with appropriate modification for the relaxed controls (definitions given below). The system (1.3) will take the following form:

$$\dot{x}^\varepsilon = \int \bar{G}(x^\varepsilon, \alpha) m_t(d\alpha) + G_0(x^\varepsilon, \xi^\varepsilon(t)) + F(x^\varepsilon, \xi^\varepsilon(t))/\varepsilon. \tag{2.1}$$

DEFINITION. An *admissible relaxed control* $m(\cdot)$ *for* (2.1) is also a measure-valued random variable (as above) but $\int_0^t f(s, \alpha) m(ds\, d\alpha)$ is progressively measurable with respect to $\{\mathcal{F}_t^\varepsilon\}$, where $\mathcal{F}_t^\varepsilon$ is the minimal σ-algebra measuring $\{\xi^\varepsilon(s), x^\varepsilon(s), s \leqq t\}$. Also, we impose $m([0, t] \times U) = t$ for all $t \geqq 0$. As in the definition given in § 1, there is also a derivative $m_t(\cdot)$, where the $m_t(B)$ are $\mathcal{F}_t^\varepsilon$-measurable for Borel B. We sometimes use the symbol $m^\varepsilon(\cdot)$ or $m_t^\varepsilon(\cdot)$ for the relaxed controls, when (2.1) is used.

The scaling in (2.1) is a common way of getting a wide bandwidth, noise-driven system [5], [13], [14]. Other forms for $\xi^\varepsilon(\cdot)$ can be used. Many examples of alternatives are in [5], where the use of perturbed test functions for weak convergence is illustrated. In particular, Example 2.2 of [5, chap. 4] describes a noise process composed of suitably scaled sums of "physical impulses"—a commonly occurring model in applications. For G_0 and F linear in ξ, noises of the "Wong-Zakai" type can also be used—although these do not represent physical models. In this paper, we use either bounded noise or Gaussian noise. For the first case (A2.1)-(A2.6) are used. The second case is covered by (A2.10). Let E_t^ε denote the expectation, conditioned on $\xi^\varepsilon(s) = \xi(s/\varepsilon^2)$, $s \leqq t$, and E_t the expectation conditioned on $\xi(s)$, $s \leqq t$.

(A2.1) $\bar{G}(\cdot, \cdot)$, $F(\cdot, \cdot)$, $G_0(\cdot, \cdot)$, $F_x(\cdot, \cdot)$ are continuous and are bounded by $O(1 + |x|)$. $G_{0,x}(\cdot, \xi)$ is continuous in x for each ξ and is bounded. $\xi(\cdot)$ is bounded, right-continuous, and $EG_0(x, \xi(t)) \to 0$, $EF(x, \xi(t)) \to 0$ as $t \to \infty$, for each x.

(A2.2) $F_{xx}(\cdot, \xi)$ is continuous for each ξ, and is bounded.

(A2.3) Let $V(x, \xi)$ denote either $\varepsilon G_0(x, \xi)$, $G_{0,x}(x, \xi)$, $F(x, \xi)$, or $F_x(x, \xi)$. Then for compact Q,

$$\varepsilon \sup_{x \in Q} \left| \int_{t/\varepsilon^2}^{\infty} E_t^{\varepsilon} V(x, \xi(s))\, ds \right| \xrightarrow{\varepsilon} 0$$

in the mean square sense, uniformly in t.

Let F_i denote the ith component of F.

(A2.4) There are continuous $\bar{F}_i(\cdot)$, $\tilde{a}(\cdot) = \{\tilde{a}_{ij}(\cdot)\}$ such that

$$\int_t^{\infty} EF'_{i,x}(x, \xi(s))F(x, \xi(t))\, ds \to \bar{F}_i(x),$$

$$\int_t^{\infty} \mathrm{E}F_i(x, \xi(s))F_j(x,(t))\, ds \to \frac{\tilde{a}_{ij}(x)}{2},$$

as $t \to \infty$, and the convergence is uniform in any bounded x-set.

Since $\tilde{a}_{ij}(\cdot)$ is not necessarily equal to $\tilde{a}_{ji}(\cdot)$ for $i \neq j$, and we need a symmetric covariance matrix below, we define $a(x) = \frac{1}{2}[\tilde{a}(x) + \tilde{a}'(x)]$.

(A2.5) For each compact set Q, and all i, j,

(a) $$\sup_{x \in Q} \varepsilon \left| \int_{t/\varepsilon}^{\infty} d\tau \int_{\tau}^{\infty} ds [E_{t/\varepsilon^2} F'_{i,x}(x, \xi(s))F(x, \xi(\tau)) - EF'_{i,x}(x, \xi(s))F(x, \xi(\tau))] \right| \to 0;$$

(b) $$\sup_{x \in Q} \varepsilon \left| \int_{t/\varepsilon^2}^{\infty} d\tau \int_{\tau}^{\infty} ds [E_{t/\varepsilon^2} F_i(x, \xi(s))F_j(x, \xi(\tau)) - EF_i(x, \xi(s))F_j(x, \xi(\tau))] \right| \to 0,$$

in the mean square sense as $\varepsilon \to 0$, uniformly in t. The last sentence also holds when the bracketed terms are replaced by their x-gradients.

Remark. If $\xi(\cdot)$ is stationary, then we need not let $t \to \infty$ in (A2.4), but can set $t = 0$. As (A2.4) is written, it allows asymptotic stationarity, i.e., it allows the effects of the initial condition to "disappear." A similar interpretation can be given for (A2.3), since the "stationary" values are $EV(x, \xi(t)) = 0$. Essentially, (A2.5b) is a condition on the rate of convergence of the conditional expectation $\int_{\tau}^{\infty} ds\, E_t F_i(x, \xi(s))F_j(x, \xi(\tau))$ to $\tilde{a}_{ij}(x)$ as $\tau - t \to \infty$, and similarly for (A2.5a). They can be shown [5, p. 82] to be satisfied if $\xi(\cdot)$ satisfied a uniform mixing condition with mixing rate $\phi(\cdot)$, where $\int_0^{\infty} \phi^{1/2}(s)\, ds < \infty$ (e.g., a finite state ergodic Markov chain) or a stable ARMA model with bounded and independently and identically distributed inputs. Very similar conditions were used in [13], [14], but where the $\xi(\cdot)$ were restricted to a class of Markov processes. The conditions in § 3 are closer to "ergodic" conditions, and are often easier to verify.

Define $\bar{b}(x, \alpha) = \bar{G}(x, \alpha) + \bar{F}(x)$ and the operators A^m (when m is a feedback relaxed control m_x), and A^{α} and A^u as follows:

$$A^{\alpha} f(x) = F'_x(x)\bar{b}(x, \alpha) + \frac{1}{2} \sum_{i,j} a_{ij}(x) f_{x_i x_j}(x),$$

$$A^m f(x) = \int A^{\alpha} f(x) m_x(d\alpha),$$

and for A^u, we replace the α in the definition of A^α by the classical control function $u(\cdot)$. For a fixed control value α, A^α will be the operator of the process that is the weak limit of $\{x^\varepsilon(\cdot)\}$.

(A2.6) The martingale problem for operator A^m has a unique solution for each relaxed admissible feedback control $m_x(\cdot)$, and each initial condition. The process is a Feller process. The solution of (2.1) is unique in the weak sense for each $\varepsilon > 0$. Also $a(x) = \sigma(x)\sigma'(x)$ for some continuous finite-dimensional matrix $\sigma(\cdot)$.

Remark. The uniqueness and existence is guaranteed if the operator A^m is that for the system

$$dx = \tilde{b}(x)\,dt + \begin{pmatrix} \int \hat{b}(x,\alpha) m_x(d\alpha)\,dt \\ 0 \end{pmatrix} + \begin{pmatrix} \sigma(x)\,dw \\ 0 \end{pmatrix}, \tag{2.2}$$

where

$$a(x) = \begin{bmatrix} \sigma_1(x)\sigma_1'(x) & 0 \\ 0 & 0 \end{bmatrix},$$

and where $\sigma_1\sigma_1' \geqq \delta I$ for all x and some $\delta > 0$, $\tilde{b}(\cdot)$ and $\sigma_1(\cdot)$ are Lipschitz-continuous and $\hat{b}(\cdot,\cdot)$ is merely bounded and Borel-measurable and the dimensions of (the vector) $\hat{b}$ and (square matrix) $\sigma_1\sigma_1'$ are equal.

Let $\mathcal{M}$ denote the space of *probability measures* on the Borel sets of $R^r \times U$, with the "weak compact" topology where $P_n \to P$ if and only if $\int f(x,\alpha) P_n(dx\,d\alpha) \to \int f(x,\alpha) P(dx\,d\alpha)$ for each continuous function $f(\cdot)$ with compact support. For an admissible relaxed control for (2.1) and (1.1), respectively, define the (occupation) measure-valued random variables $P_T^{m,\varepsilon}(\cdot)$ and $P_T^m(\cdot)$ by, respectively,

$$P_T^{m,\varepsilon}(B \times C) = \frac{1}{T}\int_0^T I_{\{x^\varepsilon(t) \in B\}} m_t(C)\,dt,$$

$$P_T^m(B \times C) = \frac{1}{T}\int_0^T I_{\{x(t) \in B\}} m_t(C)\,dt.$$

We sometimes write $m^\varepsilon(\cdot)$, if the model is (2.1). If the relaxed control for (1.1) is of the feedback form (m_x or $u(x)$), then we use the modification

$$P_T^m(B) = \frac{1}{T}\int_0^T I_{\{x(t) \in B\}}\,dt$$

(or with u replacing m), and similarly define $P_T^{m,\varepsilon}(B)$, $P_T^{u,\varepsilon}(B)$ for feedback $m(\cdot)$ and $u(\cdot)$.

Let $\{m^\varepsilon(\cdot)\}$ be a given sequence of admissible relaxed controls and let $u^\delta(\cdot)$ be defined by (A2.8) below. The value of δ is fixed in (A2.7).

(A2.7) The sets of random variables

$$\{x^\varepsilon(t), \text{ small } \varepsilon > 0,\ t \in \text{dense set in } [0,\infty),\ m^\varepsilon \text{ used}\},$$

$$\{x^\varepsilon(t), \text{ small } \varepsilon > 0,\ t \in \text{dense set in } [0,\infty),\ u^\delta \text{ used}\}$$

are tight.

Actually (A2.7) is used only because it implies that (A2.7a) holds—so we can use (A2.7a) in lieu of (A2.7).

(A2.7a) The sets of measure-valued random variables

$$\{P_T^{m^\varepsilon,\varepsilon}(\cdot), \text{ small } \varepsilon>0,\ T<\infty\},$$

$$\{P_T^{u^\delta,\varepsilon}(\cdot), \text{ small } \varepsilon>0,\ T<\infty\}$$

are tight.

In turn, (A2.7a) is implied by (A2.7b).

(A2.7b) There is a function $q(\cdot)$ such that $0\leqq q(x)\to\infty$ as $|x|\to\infty$ and

$$\overline{\lim_{\varepsilon,T}} E^{m^\varepsilon,\varepsilon}\frac{1}{T}\int_0^T q(x^\varepsilon(s))\,ds<\infty,$$

and similarly for u^δ replacing m^ε.

A criterion of a Lyapunov function type for (A2.7b) is given at the end of the section.

In many applications, a stabilizing control is used—in the sense that the system is stable for any set of bounded disturbances. Then an additional bounded control term $\bar{G}(x,u)$ is added, and we choose the new control to minimize an average cost. Of course, if the state space is compact, as for the "Lyapunov exponent" problem in § 4, then (A2.7) always holds. In lieu of a "universal stability condition," a condition on the minimum (over the control values) magnitude of the cost $k(\cdot)$ as $|x|\to\infty$ was used in [3] (for the model (1.1)) to get that an optimal control for that model is "stabilizing." Perhaps a similar idea can be used here. But this point will not be pursued.

(A2.8) For $\delta>0$, there is a continuous δ-optimal control for (1.1) and (1.2), for which the martingale problem has a unique solution for each initial condition. The solution is a Feller process and there is a unique invariant measure $\mu(u^\delta,\cdot)$. (u^δ is δ-optimal in the sense that $\bar{\gamma}(u^\delta)\leqq\bar{\gamma}(m_x)+\delta$ for any feedback relaxed control m_x for which there is a stationary solution to the associated martingale problem and the initial condition is the invariant distribution.)

(A2.9) $k(\cdot)$ is bounded and continuous.

Remark. The existence of such *smooth δ-optimal controls* (for any $\delta>0$) is dealt with in [7]. They will exist under an appropriate stability condition on the uncontrolled (1.1), and either nondegeneracy of (1.1) or for a system of the form (2.2) [7]. It turns out that $\gamma(u^\delta)=\bar{\gamma}(u^\delta)$ w.p.1 (this follows from the method of proof of Theorem 1 below, or from the method in [3], under the conditions there).

(A2.10) (Gaussian case.) $\xi(\cdot)$ is a stable Gauss-Markov process with a stationary transition function and let $F(x,\xi)=F(x)\xi$, $G_0(x,\xi)=G_0(x)\xi$, where $\bar{G}$, G_0, and F satisfy the smoothness in (A2.1)-(A2.2). Define $\bar{F}(\cdot)$ and $a(\cdot)$ as in (A2.4). (Note that all other parts of (A2.3)-(A2.5) hold.)

THEOREM 1. *Assume either* (A2.1)-(A2.9) *or* (A2.6)-(A2.10) (*with either* (A2.7) *or* (A2.7a) *or* (A2.7b) *used*). *Let* (2.1) *have a unique solution for each admissible relaxed control and each ε. Then* (1.8a) *and* (1.8b) *hold.*

Proof. We do the "Gaussian" case only. The other case is treated in essentially the same way. Let $\mathscr{D}$ be a (countable) measure determing set of bounded real-valued continuous functions on R^r having continuous second partial derivatives and compact support. Let $m_t^\varepsilon(\cdot)$ be the relaxed control in (A2.7). For a test function $f(\cdot)\in\mathscr{D}$, define the test function perturbations (the change of scale $\tau/\varepsilon^2\to\tau$ yielding the right sides

of the equations below will be used frequently and often without specific mention):

$$f_0^\varepsilon(x,t)=\int_t^\infty E_t^\varepsilon f_x'(x)G_0(x,\xi^\varepsilon(\tau))\,d\tau$$
$$=\varepsilon^2\int_{t/\varepsilon^2}^\infty E_t^\varepsilon f_x'(x)G_0(x,\xi(\tau))\,d\tau=O(\varepsilon^2)|\xi^\varepsilon(t)|,$$
$$f_1^\varepsilon(x,t)=\int_t^\infty E_t^\varepsilon f_x'(x)F(x,\xi^\varepsilon(\tau))\,d\tau/\varepsilon$$
$$=\varepsilon\int_{t/\varepsilon^2}^\infty E_t^\varepsilon f_x'(x)F(x,\xi(\tau))\,d\tau=O(\varepsilon)|\xi^\varepsilon(t)|,$$
$$f_2^\varepsilon(x,t)=\frac{1}{\varepsilon^2}\int_t^\infty d\tau\int_\tau^\infty ds\,\{E_t^\varepsilon[f_x'(x)F(x,\xi^\varepsilon(s))]_x'F(x,\xi^\varepsilon(\tau))$$
$$-E[f_x'(x)F(x,\xi^\varepsilon(s))]_x'F(x,\xi^\varepsilon(\tau))\}$$
$$=\varepsilon^2\int_{t/\varepsilon^2}^\infty d\tau\int_\tau^\infty ds\{E_t^\varepsilon[f_x'(x)F(x,\xi(s))]_x'F(x,\xi(\tau))$$
$$-E[f_x'(x)F(x,\xi(s))]_x'F(x,\xi(\tau))\}$$
$$=O(\varepsilon^2)[|\xi^\varepsilon(t)|^2+1].$$

To evaluate the integrals write $F(x,\xi)=F(x)\xi$, and use the conditional expectations or expectations associated with the Gauss-Markov process.

Define the perturbed test function

$$f^\varepsilon(t)=f(x^\varepsilon(t))+\sum_{i=0}^2 f_i^\varepsilon(x^\varepsilon(t),t).$$

The operator $\hat{A}^{m^\varepsilon,\varepsilon}$ and its domain $\mathcal{D}(\hat{A}^{m^\varepsilon,\varepsilon})$ are defined in the Appendix. By a direct calculation, using the correlation and conditional expectation properties of the Gauss-Markov process $\xi(\cdot)$, we get that $f(x^\varepsilon(\cdot))$ and the $f_i^\varepsilon(x^\varepsilon(\cdot),\cdot)$ are all in $\mathcal{D}(\hat{A}^{m,\varepsilon})$, and[1]

$$\hat{A}^{m^\varepsilon,\varepsilon}f(x^\varepsilon(t))=f_x'(x^\varepsilon(t))\dot{x}^\varepsilon(t)$$
$$=f_x'(x^\varepsilon(t))\left[\int\bar{G}(x^\varepsilon(t),\alpha)m_t^\varepsilon(d\alpha)\right.$$
$$\left.+G_0(x^\varepsilon(t),\xi^\varepsilon(t))+F(x^\varepsilon(t),\xi^\varepsilon(t))/\varepsilon\right].$$

(The "small" perturbations f_i^ε are added to the test function to facilitate the averaging of the "noise" in the above expression. Related calculations are used in [13] and [14].) continuing, we have

$$\hat{A}^{m^\varepsilon,\varepsilon}f_0^\varepsilon(x^\varepsilon(t),t)=-f_x'(x^\varepsilon(t))G_0(x^\varepsilon(t),\xi^\varepsilon(t))$$
$$+\int_t^\infty[E_t^\varepsilon f_x'(x^\varepsilon(t))G_0(x^\varepsilon(t),\xi^\varepsilon(s))]_x'\dot{x}^\varepsilon(t)\,ds/\varepsilon,$$
$$\hat{A}^{m^\varepsilon,\varepsilon}f_1^\varepsilon(x^\varepsilon(t),t)=-f_x'(x^\varepsilon(t))F(x^\varepsilon(t),\xi^\varepsilon(t))/\varepsilon$$
$$+\int_t^\infty ds[E_t^\varepsilon f_x'(x^\varepsilon(t))F(x^\varepsilon(t),\xi^\varepsilon(s))]_x'\dot{x}^\varepsilon(t)/\varepsilon,$$

[1] Acting on functions $f(x^\varepsilon(t))$, the operator $\hat{A}^{m^\varepsilon,\varepsilon}$ is just a differentiation operator. It is also a differentiation operator when acting on functions such as $f_1^\varepsilon(x^\varepsilon(t),t)$ or $f_2^\varepsilon(x^\varepsilon(t),t)$. The calculation is actually a differentiation with respect to the t appearing in $x^\varepsilon(t)$ and also in the lower limit of integration. The t in E_t^ε plays no role in the calculation, owing to the way that E_t^ε appears in the definition of $\hat{A}^{m^\varepsilon,\varepsilon}$.

$$\hat{A}^{m^\varepsilon,\varepsilon}f_2^\varepsilon(x^\varepsilon(t),t)=-\frac{1}{\varepsilon^2}\int_t^\infty ds\{E_t^\varepsilon[f_x'(x^\varepsilon(t))F(x^\varepsilon(t),\xi^\varepsilon(s))]_x'F(x^\varepsilon(t),\xi^\varepsilon(t))$$
$$-E[f_x'(x)F(x,\xi^\varepsilon(s))]_x'F(x,\xi^\varepsilon(t))|_{x=x^\varepsilon(t)}\}$$
$$+[f_2^\varepsilon(x,t)]_x'\dot{x}^\varepsilon(t)|_{x=x^\varepsilon(t)}.$$

The dominant component of the right-hand term of the last expression is $O(\varepsilon)[1+|\xi^\varepsilon(t)|^3]$. This can be seen by using the double scale change in $f_2^\varepsilon(x,t)$ (namely, $s/\varepsilon^2\to s$, $\tau/\varepsilon^2\to\tau$) and the properties of the Gauss-Markov process $\xi(\cdot)$.

We have

$$|f(x^\varepsilon(t))-f^\varepsilon(t)|=O(\varepsilon)[|\xi^\varepsilon(t)|^2+1]. \tag{2.3a}$$

By adding $\sum_{i=0}^2\hat{A}^{m^\varepsilon,\varepsilon}f_i^\varepsilon(t)$ to $\hat{A}^{m^\varepsilon,\varepsilon}f(x^\varepsilon(t))$, subtracting from $A^{m^\varepsilon}f(x^\varepsilon(t))=\int A^\alpha f(x^\varepsilon(t))m_t^\varepsilon(d\alpha)$ and canceling terms where possible, we get

$$|\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(t)-A^{m^\varepsilon}f(x^\varepsilon(t))|=O(\varepsilon)[|\xi^\varepsilon(t)|^3+1]. \tag{2.3b}$$

(The perturbations were constructed just to get the cancellations of the bad terms or the replacement by their averages.) All the $O(\varepsilon)$ are uniform in t, ε, and ω. By (6.4) of the Appendix (with our f^ε replacing the q there), the function

$$M_f^\varepsilon(t)=f^\varepsilon(t)-f^\varepsilon(0)-\int_0^t\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(s)\,ds \tag{2.4}$$

is a zero mean martingale. We next show that $M_f^\varepsilon(t)/t\xrightarrow{P}0$ as $t\to\infty$ and $\varepsilon\to0$ in any way at all.

Write (where $[t]$ denotes the greatest integer part of t)

$$\frac{M_f^\varepsilon(t)}{t}=\frac{1}{t}[(M_f^\varepsilon(t)-M_f^\varepsilon([t]))+M_f^\varepsilon(0)]+\frac{1}{t}\sum_{n=0}^{[t]-1}[M_f^\varepsilon(n+1)-M_f^\varepsilon(n)]. \tag{2.5}$$

Using the fact that $f(\cdot)$ is bounded and (2.3), (2.5), and the martingale property of $M_f^\varepsilon(\cdot)$, we get that $E[M_f^\varepsilon(t)/t]^2=O(1)/t$. The fact that $M_f^\varepsilon(t)/t$, $f^\varepsilon(t)/t$, and $f^\varepsilon(0)/t$ all go to zero in probability as $t\to\infty$ (uniformly in ε) together with (2.4) and the second line of (2.3) implies that as $t\to\infty$ and $\varepsilon\to0$,

$$\int_0^t A^{m^\varepsilon}f(x^\varepsilon(s))\,ds/t\xrightarrow{P}0. \tag{2.6a}$$

By the definition of $P_T^{m^\varepsilon,\varepsilon}(\cdot)$, (2.6a) can be written as

$$\int A^\alpha f(x)P_T^{m^\varepsilon,\varepsilon}(dx\,d\alpha)\xrightarrow{P}0\quad\text{as } T\to\infty\text{ and }\varepsilon\to0. \tag{2.6b}$$

Now, let the control be the classical control function $u^\delta(\cdot)$, and choose a weakly convergent subsequence of the set of random variables $\{P_T^{u^\delta,\varepsilon}(\cdot),\varepsilon,T\}$ (and also such that $1/t\int_0^t A^{u^\delta}f(x^\varepsilon(s))\,ds\to0$ w.p.1 for all $f(\cdot)\in\mathcal{D}$), indexed by ε_n, T_n, and with (random) limit denoted by $\tilde{\mu}(\cdot)$. We let the limits $\tilde{\mu}(\cdot)$ be defined on some probability space $(\bar{\Omega},\bar{P},\bar{\mathcal{F}})$ with generic variable $\bar{\omega}$. Now, (2.6b) implies that

$$\int A^{u^\delta}f(x)\tilde{\mu}(dx)=0\quad\text{for }\bar{P}\text{-almost all }\bar{\omega}. \tag{2.7}$$

Since our class of $f(\cdot)$ is measure-determining, (2.7) implies that almost all realizations of $\tilde{\mu}(\cdot)$ are invariant measures for (1.1) (under u^δ). (This is proved by a slight extension of Proposition 9.2 of [8].) By uniqueness of the invariant measure,

we can take $\mu(u^\delta, \cdot) = \tilde{\mu}(\cdot)$ for all $\bar{\omega}$, and the limit $\tilde{\mu}(\cdot)$ does not depend on the chosen subsequence ε_n, T_n. Furthermore, by the definition of $P_T^{u^\delta,\varepsilon}(\cdot)$,

$$\int_0^t k(x^\varepsilon(s), u^\delta(x^\varepsilon(s)))\, ds/t = \int_0^t k(x, u^\delta(x)) P_t^{u^\delta,\varepsilon}(dx)$$

$$\xrightarrow{P} \int k(x, u^\delta(x))\mu(u^\delta, dx) = \bar{\gamma}(u^\delta).$$

Next, choose a weakly convergent subsequence of $\{P_T^{m^\varepsilon,\varepsilon}(\cdot), \varepsilon, T\}$ (and also such that (2.6a) $\to 0$ w.p.1 for all $f(\cdot) \in \mathscr{D}$) indexed by ε_n, T_n, and with limit denoted by $\tilde{P}(\cdot)$ (again, defined on some probability space $(\bar{\Omega}, \bar{P}, \bar{\mathscr{F}})$). For each $\bar{\omega}$, we can factor $\tilde{P}(\cdot)$ as $\tilde{P}(dx\, d\alpha) = m_x(d\alpha)\mu(dx)$. We can suppose that the $m_x(B)$ are x-measurable for each Borel B and $\bar{\omega}$.

By (2.6), for all $f(\cdot) \in \mathscr{D}$,

$$\iint A^\alpha f(x) m_x(d\alpha)\mu(dx) = 0 \quad \text{for } \bar{P}\text{-almost all } \bar{\omega}. \tag{2.8}$$

This implies that (for $\bar{\omega}$ almost all $\bar{\omega}$), $\mu(\cdot)$ is an invariant measure for the process (1.1) with relaxed feedback control $m_x(\cdot)$. As above we also have

$$\iint k(x, \alpha) m_x(d\alpha)\mu(dx) = \lim_{\varepsilon_n, T_n} \gamma_{T_n}^{\varepsilon_n}(m^\varepsilon) = \bar{\gamma}(m_x). \tag{2.9}$$

But, by the δ-optimality of $u^\delta(\cdot)$, for almost all $\bar{\omega}$ we have $\bar{\gamma}(m_x) \geqq \bar{\gamma}(u^\delta) - \delta$. Since this is true for all the limits of the tight set $\{P_T^{m^\varepsilon,\varepsilon}(\cdot); \varepsilon, T\}$, (1.8b) follows. □

A Lyapunov function criterion for (A2.7b). For illustrative purposes, we do a simple case where the stabilizing dynamics "$\bar{G}$" are dominant. In [5, Chap. 6.6], a related calculation is given for a case where the F-terms must be taken into account. Our case concerns the situation where a globally stabilizing control for the limit system is given a priori, and the control to be chosen is bounded. The main problem arises from the $F(x, \xi)/\varepsilon$ term. We will use the following assumptions.

(A2.11) There is a twice continuously differentiable function $0 \leqq V(x)$ and a $0 \leqq g(\cdot)$ and $K < \infty$ such that $g(x) \to \infty$ as $|x| \to \infty$ and $V_x'(x)\bar{G}(x, \alpha) \leqq -g(x) + K$, all x, α. Also $V_x(x)/g(x) \to 0$ as $|x| \to \infty$.

(A2.12) $G_0(\cdot, \cdot)$, $F(\cdot, \cdot)$ are bounded and continuous, $F(\cdot, \xi)$ has a bounded and continuous (uniformly in ξ) derivative, and (2.1) has a unique solution for each admissible relaxed control. $\xi(\cdot)$ is right-continuous and bounded $\bar{G}(\cdot, \cdot)$ is continuous and is bounded by $O(1 + |x|)$.

Define the perturbation to $V(\cdot)$:

$$V_1^\varepsilon(x, t) = \frac{1}{\varepsilon} \int_t^\infty E_t^\varepsilon V_x'(x) F(x, \xi^\varepsilon(s))\, ds = \varepsilon \int_{t/\varepsilon^2}^\infty E_{t/\varepsilon^2} V_x'(x) F(x, \xi(s))\, ds.$$

(A2.13) The limits below are uniform in t, ω (w.p.1),

$$\lim_{|x| \to \infty} |V_1^\varepsilon(x, t)/V(x)| = 0, \qquad \overline{\lim_{|x| \to \infty}} |(V_1^\varepsilon(x, t))_x / g(x)| = O(\varepsilon),$$

$$\overline{\lim_{|x| \to \infty}} |(V_1^\varepsilon(x, t))_x' \bar{G}(x, \alpha)/g(x)| = O(\varepsilon).$$

THEOREM 2. *Assumption* (A2.7b) *holds under* (A2.11)-(A2.13), *for each initial condition* $x^\varepsilon(0) = x$.

Proof. Fix $\{m^\varepsilon\}$ and define the perturbed Lyapunov function $V^\varepsilon(x^\varepsilon(t), t) = V(x^\varepsilon(t)) + V_1^\varepsilon(x^\varepsilon(t), t)$. Similarly to what has been done in Theorem 1, $V^\varepsilon(x^\varepsilon(t), t) \in \mathscr{D}(\hat{A}^{m^\varepsilon,\varepsilon})$ and

$$\begin{aligned}\hat{A}^{m^\varepsilon,\varepsilon}V^\varepsilon(x^\varepsilon(t), t) &= V_x'(x^\varepsilon(t))\int \bar{G}(x^\varepsilon(t), \alpha)m_t^\varepsilon(d\alpha)\\ &\quad + V_x'(x^\varepsilon(t))G_0(x^\varepsilon(t), \xi^\varepsilon(t))\\ &\quad + \frac{1}{\varepsilon}\int_t^\infty ds\, E_t^\varepsilon[V_x'(x^\varepsilon(t))F(x^\varepsilon(t), \xi^\varepsilon(s))]_x'\\ &\quad \cdot\left[G_0(x^\varepsilon(t), \xi^\varepsilon(t)) + \int \bar{G}(x^\varepsilon(t), \alpha)m_t^\varepsilon(d\alpha) + F(x^\varepsilon(t), \xi^\varepsilon(t))/\varepsilon\right].\end{aligned} \tag{2.10}$$

By (A2.13) and a change of scale $s/\varepsilon^2 \to s$ we have that the right-hand term of (2.10) is bounded in absolute value by $O(\varepsilon)g(x) + o(g(x))$, where $o(g)/g \to 0$ as $g \to \infty$. Thus there is $K_1 < \infty$ such that for small ε

$$\hat{A}^{m^\varepsilon,\varepsilon}V^\varepsilon(x^\varepsilon(t), t) \leqq -\tfrac{1}{2}g(x^\varepsilon(t)) + K_1.$$

Hence

$$\begin{aligned}&[E^{m,\varepsilon}V(x^\varepsilon(t)) + E^{m,\varepsilon}V_1^\varepsilon(x^\varepsilon(t), t) - V(x) - V_1^\varepsilon(x, 0)]/t\\ &\leqq -\frac{1}{2t}E^{m^\varepsilon,\varepsilon}\int_0^t g(x^\varepsilon(s))\, ds + K_1.\end{aligned} \tag{2.11}$$

By the first line of (A2.13) and the fact that $V(x) \geqq 0$, by taking $t \to \infty$ we get

$$\overline{\lim_{\varepsilon,t}}\frac{1}{t}E^{m^\varepsilon,\varepsilon}\int_0^t g(x^\varepsilon(s))\, ds \leqq 2K_1,$$

which is (A2.7b). □

3. Alternative conditions. In this section we redo Theorem 1 under somewhat different conditions. The perturbed test function is only "first-order" here and (2.3) will not hold. But similar results are obtained via a direct averaging method of the type introduced in [5, chap. 5]. We will use either bounded "mixing" or Gaussian noise, as in § 2, and subsets of the following conditions. Let E_t denote the expectation given $\xi(s)$, $s \leqq t$.

(A3.1) $\xi(\cdot)$ is bounded, and right-continuous $G_0(\cdot,\cdot)$, $\bar{G}(\cdot,\cdot)$, $F(\cdot,\cdot)$, and $F_x(\cdot,\cdot)$ are continuous and bounded by $O(1+|x|)$.

(A3.2) $\int_t^\infty E_sF(x, \xi(s))\, ds$, $\int_t^\infty E_t[f_x'(x)F(x, \xi(s))]_x'F(x, \xi(t))\, ds$, are bounded and x-continuous uniformly on each compact x-set and uniformly in t, ω.

(A3.3) $1/T\int_t^{t+T} E_tG_0(x, \xi(s))\, ds \xrightarrow{P} 0$, for each x as t and $T \to \infty$.

(A3.4) There are continuous $\bar{F}(\cdot)$, $a(\cdot)$ such that with A_0 (acting on twice continuously differentiable real-valued functions $f(\cdot)$ with compact support) given by

$$A_0f(x) = f_x'(x)\bar{F}(x) + \frac{1}{2}\sum_{i,j} a_{ij}(x)f_{x_ix_j}(x), \qquad a_{ij}(x) = a_{ji}(x),$$

we have

$$\frac{1}{T}\int_t^{t+T} ds \int_s^\infty du\, E_t[f_x'(x)F(x, \xi(u))]_x'F(x, \xi(s)) \xrightarrow{P} A_0f(x),$$

for each x as t and $T \to \infty$.

(A3.5) $\xi(\cdot)$ is a stable Gauss-Markov process, with a stationary transition function, and $F(x,\xi)=F(x)\xi$, $G_0(x,\xi)=G_0(x)\xi$, and $F(\cdot)$, $\bar{G}(\cdot,\cdot)$ and $G_0(\cdot)$ have the smoothness of (A3.1). (We continue to define $\bar{F}(\cdot)$, $a(\cdot)$, and A_0 as in (A3.4), when (A3.5) is used.)

As in § 1, set $A^\alpha f(x)=f_x'(x)\bar{G}(x,\alpha)+A_0 f(x)$, and $\bar{b}(x,\alpha)=\bar{G}(x,\alpha)+\bar{F}(x)$.

THEOREM 3. *Assume* (A2.6)-(A2.9) *with either* (A2.7) *or* (A2.7a) *or* (A2.7b) *used, and either* (A3.1)-(A3.4) *or else* (A3.5). *Let* (2.1) *have a unique solution for each admissible relaxed control and each* $\varepsilon>0$. Then (1.8a) *and* (1.8b) *hold.*

Proof. Let $f(\cdot)$ be as in Theorem 1. We use the "direct averaging first-order perturbed test function method" of [5, Chap. 5], [9], [1], but the development here is self-contained. Define $f_1^\varepsilon(x,t)$ as in Theorem 1 and set $f^\varepsilon(t)=f(x^\varepsilon(t))+f_1^\varepsilon(x^\varepsilon(t),t)$. Then (write x for $x^\varepsilon(t)$ for convenience here) $f^\varepsilon(\cdot)\in\mathcal{D}(\hat{A}^{m^\varepsilon,\varepsilon})$ and

$$\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(t)=f_x'(x)\left[\int\bar{G}(x,\alpha)m_t^\varepsilon(d\alpha)+G_0(x,\xi^\varepsilon(t))\right]$$
$$+\frac{1}{\varepsilon^2}\int_t^\infty ds[E_t^\varepsilon f_x'(x)F(x,\xi^\varepsilon(s))]_x'F(x,\xi^\varepsilon(t))$$
$$+\text{terms of order } O(\varepsilon)[|\xi^\varepsilon(t)|^2+1].$$

(See the expressions given above (2.3).) When we use the scale change $s/\varepsilon^2\to s$, the second term can be seen to be bounded in mean square for the bounded noise case and $O(1)[|\xi^\varepsilon(t)|^2+1]$ in the Gaussian case.

Define the martingale

$$M_f^\varepsilon(t)=f^\varepsilon(t)-f^\varepsilon(0)-\int_0^t\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(s)\,ds.$$

If

$$M_f^\varepsilon(t)/t\xrightarrow{P}0\quad\text{as }\varepsilon\to0,\quad t\to\infty, \tag{3.1}$$

then as in Theorem 1, we have

$$\int_0^t\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(s)\,ds/t\xrightarrow{P}0\quad\text{as }\varepsilon\to0,\quad t\to\infty.$$

If we also have that

$$\frac{1}{t}\int_0^t[\hat{A}^{m^\varepsilon,\varepsilon}f^\varepsilon(s)-A^{m^\varepsilon}f(x^\varepsilon(s))]\,ds\xrightarrow{P}0\quad\text{as }\varepsilon\to0,\quad t\to\infty, \tag{3.2}$$

(and also for u^δ used in lieu of $m^\varepsilon(\cdot)$), then the proof can be completed as in Theorem 1. Thus, we need only show (3.1) and (3.2).

To get (3.1), we use the representation (2.5). The martingale difference $M_f^\varepsilon(n+1)-M_f^\varepsilon(n)$ equals

$$f^\varepsilon(n+1)-f^\varepsilon(n)-\int_n^{n+1}ds\left[f_x'(x^\varepsilon(s))\int\bar{G}(x^\varepsilon(s),\alpha)m_s^\varepsilon(d\alpha)+G_0(x^\varepsilon(s),\xi^\varepsilon(s))\right]$$
$$+\int_n^{n+1}ds\,O(1)[|\xi^\varepsilon(s)|^2+1]. \tag{3.3}$$

Since the mean square value of (3.3) is bounded uniformly in n, ω, ε, we get that $E[M_f^\varepsilon(t)]^2/t=O(1/t)$ and (3.1) holds, exactly as for Theorem 1.

We now prove (3.2). To *simplify the proof, we drop the terms* $\int \bar{G}(x,\alpha)m_t^\varepsilon(d\alpha)$ *and* $G_0(x,\xi)$. The first dropped term causes no problems (as in Theorem 1) and the second is dealt with by an averaging method similar to that employed below. Now, we have

$$\begin{aligned}
&\frac{1}{t}\int_0^t \hat{A}^{m^\varepsilon,\varepsilon} f^\varepsilon(s)\,ds \\
&= \frac{1}{t}\int_0^t ds \int_s^\infty du\, E_s^\varepsilon [f_x'(x^\varepsilon(s))F(x^\varepsilon(s),\xi^\varepsilon(u))]_x' F(x,\xi^\varepsilon(s))/\varepsilon^2 \\
&\qquad + \text{negligible terms} \\
&= \frac{\varepsilon^2}{t}\int_0^{t/\varepsilon^2} ds \int_s^\infty du\, E_s [f_x'(x^\varepsilon(\varepsilon^2 s)F(x^\varepsilon(\varepsilon^2 s),\xi(u))]_x' F(x^\varepsilon(\varepsilon^2 s),\xi(s)) \\
&\qquad + \text{negligible terms}
\end{aligned} \tag{3.4}$$

where the negligible terms go to zero in the mean square sense as $\varepsilon \to 0$. Henceforth, for simplicity, we consider the *scalar* case and work with only the term $f_{xx}(x)F(x,\xi(u))F(x,\xi(s))$ in (3.4). Write $t = N\Delta$ for integer N and $\Delta > 0$. Define

$$Q^\varepsilon(x,s) = \int_s^\infty du\, E_s f_{xx}(x)F(x,\xi(u))F(x,\xi(s)).$$

Then the desired term in (3.4) can be written as follows:

$$\begin{aligned}
&\frac{1}{N}\sum_1^N \frac{\varepsilon^2}{\Delta}\int_{i\Delta/\varepsilon^2}^{(i\Delta+\Delta)/\varepsilon^2} ds[E_{i\Delta}^\varepsilon Q^\varepsilon(x^\varepsilon(\varepsilon^2 s),s) - Q^\varepsilon(x^\varepsilon(\varepsilon^2 s),s)] \\
&\qquad + \frac{1}{N}\sum_1^N \frac{\varepsilon^2}{\Delta}\int_{i\Delta/\varepsilon^2}^{(i\Delta+\Delta)/\varepsilon^2} E_{i\Delta}^\varepsilon Q^\varepsilon(x^\varepsilon(\varepsilon^2 s),s)\,ds.
\end{aligned} \tag{3.5}$$

Since $E|E_{i\Delta}^\varepsilon Q^\varepsilon(x^\varepsilon(\varepsilon^2 s),s) - Q^\varepsilon(x^\varepsilon(\varepsilon^2 s),s)|^2$ is bounded uniformly in s, ε, and Δ, the first set of *summands* in (3.5) are martingale differences with uniformly (in ε, N, t) bounded mean square values. Thus the first sum is $O(1/N)$ and goes to zero in probability as $N \to \infty$, uniformly in ε, t. Let $f(x) = 0$ for $|x| \geqq K$. By [5, Thm. 4, Chap. 3], and the uniform integrability of $\{\hat{A}^{m^\varepsilon,\varepsilon} f^\varepsilon(t), \varepsilon > 0, t < \infty\}$, the sequence

$$\{[x^\varepsilon(i\Delta+\cdot) - x^\varepsilon(i\Delta)]I_{\{|x^\varepsilon(i\Delta)|\leqq 2K\}},\ i, \Delta > 0,\ \varepsilon > 0\}$$

is tight in $D[0,\infty)$ (Skorokhod topology) and has continuous limits w.p.1. Because of this, we can replace the $x^\varepsilon(\varepsilon^2 s)$ in the ith summand of the second term in (3.5) by $x^\varepsilon(i\Delta)$ for all i, and only alter the sum by an amount that goes to zero in probability (uniformly in ε and N) as $\Delta \to 0$.

Doing this replacement and using either the Gaussian property (A3.5) or else (A3.4) for the bounded noise case, and the continuity of $F(\cdot,\xi)$ (uniform in ξ in the bounded noise case) and the continuity and compact support of $f_{xx}(\cdot)$ yields that the second sum in (3.5) and

$$\frac{1}{N}\sum_1^N \frac{\varepsilon^2}{\Delta}\int_{i\Delta/\varepsilon^2}^{(i\Delta+\Delta)/\varepsilon^2} ds\, f_{xx}(x^\varepsilon(i\Delta))a(x^\varepsilon(i\Delta))/2 \tag{3.6}$$

have the same limit in probability as $N \to \infty$, $\Delta \to 0$, $\varepsilon \to 0$, $N\Delta \to \infty$. We next use the rightness of

$$\{[x^\varepsilon(i\Delta+\cdot) - x^\varepsilon(i\Delta)]I_{\{|x^\varepsilon(i\Delta)|\leqq 2K\}},\ i, \Delta > 0,\ \varepsilon > 0\}$$

again to replace the $x^\varepsilon(i\Delta)$ in (3.6) by $x^\varepsilon(\varepsilon^2 s)$, and get the same result, namely that the limit in probability is the same as $N\to\infty$, $\Delta\to 0$, $\varepsilon\to 0$, $N\Delta\to\infty$. Finally, repeating the approximation procedure used from (3.5) on for the various neglected terms yields (3.2). □

4. Extensions.

Discrete time problem. There are direct extensions to the discrete parameter model

$$X^\varepsilon_{n+1} = X^\varepsilon_n + \varepsilon \int \bar{G}(X^\varepsilon_n, \alpha) m_n(d\alpha) + \varepsilon G_0(X^\varepsilon_n, \xi^\varepsilon_n) + \sqrt{\varepsilon} F(X^\varepsilon_n, \xi^\varepsilon_n). \tag{4.1}$$

In both (4.1) and (2.1), we can allow some "state dependence" of the noise (cf. the "Markov"-dependent type used in [5, Chaps. 4.4 or 5.5]).

Approximate nonlinear filtering. In the following two applications, there is no control. In § 7 of [10], an "approximate" nonlinear filtering problem was dealt with, where the system driving and observation noises were wideband. It was shown (under a condition concerning the uniqueness of a certain invariant measure) that the average error (using the notation of that paper)

$$\lim_\varepsilon \frac{1}{T}\int_0^T E[\phi(x^\varepsilon(t)) - (P^\varepsilon(t), \phi)]^2\, dt \tag{4.2}$$

converged to what we would get if the true optimal filter were used on the "limit" process. Here $x^\varepsilon(\cdot)$ is the state of the "signal system" (say, of the form (2.1)), $\phi(\cdot)$ is bounded and continuous, and $P^\varepsilon(\cdot)$ is the measure-valued output (not necessarily the conditional distribution) of the "approximate" filters used in [12]. Via the technique of this paper, similar results can be obtained if the E in (4.2) were dropped. This is useful, since we would normally filter only one path—over a long time—and the use of the expectation might give an inappropriate measure of the filter performance.

Lyapunov exponents for wide bandwidth noise-driven systems. The theory of Lyapunov exponents is well developed for systems of the form

$$dx = Ax\, dt + \sum_{i=1}^k B_i x \circ dw_i, \tag{4.3}$$

where the $\circ$ denotes that the stochastic integral is in the "Stratonovich" sense and where the $w_i(\cdot)$ are real-valued and mutually independent standard Wiener processes [11]. The "Stratonovich" sense integral is used to be consistent with the usage in [11] and because it simplifies the identification of the limit process and its "projection" below in this case. Of practical interest are the convergence properties of numerical methods of evaluating these exponents, as well as the study of the asymptotic behavior of wideband noise-driven systems (4.4) via

$$\dot{x}^\varepsilon = Ax^\varepsilon + \sum_{i=1}^k B_i x \xi^\varepsilon_i, \tag{4.4}$$

the method of Lyapunov exponents. In (4.4), the $\xi^\varepsilon_i(\cdot)$ are orthogonal and scalar-valued processes. Of particular interest is whether the exponents for (4.4) converge to those for the limit system (which will be of the general form of (4.3)) as $\varepsilon\to 0$.

Under the conditions of Theorem 3 on $\xi^\varepsilon_i(\cdot) = \xi(\cdot/\varepsilon^2)$, the above orthogonality condition, and the normalization

$$\frac{1}{T}\int_t^{t+T} ds \int_s^\infty E_t \xi_i(s)\xi_i(u)\, du \to \frac{1}{2}$$

in probability as t and T go to ∞, the $x(\cdot)$ of (4.3) is the weak limit of (4.4), if the initial conditions converge. We can assume this normalization to hold in general, since otherwise we absorb the "constants" into the B_i in the obvious way.

Define $y = x/|x|$. Then

$$\dot{y} = \dot{x}/|x| - x[x'\dot{x}]/|x|^{3/2}$$

and

$$\dot{y}^\varepsilon = Ay^\varepsilon + \sum_{i=1}^{k} B_i y^\varepsilon \xi_i^\varepsilon - y^\varepsilon [y^{\varepsilon\prime} A y^\varepsilon] - y^\varepsilon \left[y^{\varepsilon\prime} \sum_{i=1}^{k} \xi_i^\varepsilon B_i y^\varepsilon \right]. \tag{4.5}$$

Assume the noise conditions of Theorem 3. Then, it is not hard to show that $P\{x^\varepsilon(s) \neq 0, \text{ any } s \leqq T\} = 1$ for all ε, T.

Of interest is the calculation of quantities such as $\lim_t E \int_0^t q(y^\varepsilon(s))\, ds/t$ for bounded and continuous $q(\cdot)$. In the Monte Carlo evaluation of the limit, we often use

$$\frac{1}{t} \int_0^t q(y^\varepsilon(s))\, ds \tag{4.6}$$

for large t and some small ε; it is of interest to know whether or not the convergence is to the correct limit and whether it is uniform in ε and t in the sense of (1.8a). (An alternative is of course to fix $T < \infty$ and approximate $E \int_0^T q(y^\varepsilon(s))\, ds/T$ for small ε by taking many independent runs and averaging. But, the "uniformity" questions still arise.)

Define $y(t) = x(t)/|x(t)|$ and

$$q(y) = y'Ay + \frac{1}{2} \sum_{i=1}^{k} [y'(B_i + B_1')B_i y - (y'B_i y)^2],$$

and assume that $y(\cdot)$ has a unique invariant measure on the sphere (this is true under a Lie algebraic condition on the set $(A, B_i, i \leqq k)$ [11]). Then [11] the (maximal) Lyapunov exponent is the limit (which is a constant w.p.1)

$$\lim_t \int_0^t q(y(s))\, ds/t. \tag{4.7}$$

We are interested in whether or not (4.6) converges to (4.7) as $\varepsilon \to 0$ and $t \to \infty$.

By Theorem 3 $(x^\varepsilon(\cdot), y^\varepsilon(\cdot)) \Rightarrow (x(\cdot), y(\cdot))$ (Skorokhod topology), and the weak limit process $y(\cdot)$ is characterized completely by the correlation functions of the $\xi_i(\cdot)$. Let $\mu(\cdot)$ denote the assumed unique invariant measure for $y(\cdot)$. Then

$$\frac{1}{t} \int_0^t q(y^\varepsilon(s))\, ds \xrightarrow{P} \int q(y)\mu(dy) \quad \text{as } \varepsilon \to 0, \quad t \to \infty, \tag{4.8}$$

and the limit value is just the (maximum) Lyapunov exponent for $x(\cdot)$. The general method is applicable to a wide variety of noise processes and can readily be extended to yield convergence of various numerical approximations to the (maximal) Lyapunov exponent for (4.3), via use of either a discrete time approximation to (4.3) or the various interpolations that can be used to approximate the stochastic integrals.

5. Convergence of pathwise discounted costs to the ergodic cost. In this section, we treat the discounted cost result (1.9). Again, the exact sense in which the $m^\varepsilon(\cdot)$ are δ_1-optimal is left a little vague. Since $u^\delta(\cdot)$ is asymptotically δ-optimal, no matter

what the $m^\varepsilon(\cdot)$ are, the pathwise costs are (for small β, ε) no better (modulo 2δ) than the costs for the $u^\delta(\cdot)$, with an arbitrary large probability.

THEOREM 4. *Under the conditions of either Theorem* 1 *or* 2, *the limits* (1.9) *hold.*

Remarks on the Proof. The proof is essentially the same as those of Theorems 1 or 3, and we only remark on the differences. We use the discounted occupation measures

$$
\begin{aligned}
P_\beta^{m,\varepsilon}(B\times C) &= \beta \int_0^\infty e^{-\beta t} I_{\{x^\varepsilon(t)\in B\}} m_t(C)\, dt,\\
P_\beta^{m}(B\times C) &= \beta \int_0^\infty e^{-\beta t} I_{\{x(t)\in B\}} m_t(C)\, dt
\end{aligned}
\tag{5.1}
$$

and analogously for the feedback control cases.

Then the cost can be written as

$$
V_\beta^\varepsilon(m^\varepsilon) = \int\int k(x,\alpha) P_\beta^{m^\varepsilon,\varepsilon}(dx\, d\alpha).
$$

By the tightness condition (A2.7), or (A2.7a), or (A2.7b), the $\{P_\beta^{m^\varepsilon,\varepsilon}(\cdot)\}$ and $\{P^{u^\delta,\varepsilon}(\cdot)\}$ are tight. Define

$$
f_\beta^\varepsilon(t) = \beta\, e^{-\beta t} f^\varepsilon(t). \tag{5.2}
$$

This will be used in lieu of the $f^\varepsilon(\cdot)$ defined in either Theorems 1 or 3. We have

$$
\hat{A}^{m^\varepsilon,\varepsilon} f_\beta^\varepsilon(t) = -\beta^2 e^{-\beta t} f^\varepsilon(t) + \beta\, e^{-\beta t} \hat{A}^{m^\varepsilon,\varepsilon} f^\varepsilon(t). \tag{5.3}
$$

Define the martingale

$$
\begin{aligned}
&f_\beta^\varepsilon(t) - f_\beta^\varepsilon(0) - \int_0^t \hat{A}^{m^\varepsilon,\varepsilon} f_\beta^\varepsilon(s)\, ds\\
&\quad = \beta\, e^{-\beta t} f^\varepsilon(t) - \beta f^\varepsilon(0) - \int_0^t [-\beta^2 e^{-\beta s} f^\varepsilon(s) + \beta\, e^{-\beta s} \hat{A}^{m^\varepsilon,\varepsilon} f^\varepsilon(s)]\, ds.
\end{aligned}
$$

As in Theorems 1 or 3

$$
0 = \lim_{\substack{(\beta,\varepsilon)\to 0\\ t\to\infty}} \beta \int_0^t e^{-\beta s} A^{m^\varepsilon,\varepsilon} f(x^\varepsilon(s))\, ds. \tag{5.4}
$$

Thus,

$$
0 = \lim_{(\beta,\varepsilon)\to 0} \int\int A^\alpha f(x) P_\beta^{m^\varepsilon,\varepsilon}(dx\, d\alpha). \tag{5.5}
$$

Again we choose weakly convergent subsequences of the $\{P_\beta^{m^\varepsilon,\varepsilon}(\cdot)\}$ or $\{P_\beta^{u^\delta,\varepsilon}(\cdot)\}$ and continue as in the proofs of either Theorems 1 or 3 to get Theorem 4.

6. Appendix.

DEFINITION. Let $q(\cdot)$ be progressively measurable with respect to $\{\mathcal{F}_t^\varepsilon\}$, the minimal σ-algebra measuring $\{\xi^\varepsilon(s), x^\varepsilon(s), s \leqq t\}$. Suppose that there is a progressively measurable (with respect to $\{\mathcal{F}_t^\varepsilon\}$) $g(\cdot)$ such that

$$
\sup_{t\leqq T} E|g(t)| < \infty, \quad E|g(t+s) - g(t)| \to 0 \quad \text{as } s\downarrow 0 \quad \text{almost all } t, \tag{6.1}
$$

$$
\sup_{\substack{t\leqq T\\ \delta>0}} E\left|\frac{E_t^\varepsilon q(t+\delta) - q(t)}{\delta} - g(t)\right| < \infty \tag{6.2}
$$

$$
\lim_{\delta\downarrow 0} E\left|\frac{E_t^\varepsilon q(t+\delta) - q(t)}{\delta} - g(t)\right| \to 0 \quad \text{almost all } t. \tag{6.3}
$$

Then we say that $q(\cdot) \in \mathcal{D}(\hat{A}^{m,\varepsilon})$, the domain of the operator $\hat{A}^{m,\varepsilon}$ and that $\hat{A}^{m,\varepsilon} q = g$. If $q(\cdot) \in \mathcal{D}(\hat{A}^{m,\varepsilon})$, then [3, Chap. 3], [12],

$$q(t) - \int_0^t \hat{A}^{m,\varepsilon} q(s)\, ds \tag{6.4}$$

is an $\mathcal{F}_t^\varepsilon$-martingale. This martingale property will be heavily used in the proofs. We define $\hat{A}^{\alpha,\varepsilon}$ to be $\hat{A}^{m,\varepsilon}$ with m_t concentrated at α, and $\hat{A}^{u,\varepsilon}$ is defined in the obvious way

The form given for $\hat{A}^{m,\varepsilon}$ in Theorem 1 satisfies (6.1)–(6.3) if $\int \bar{G}(x, \alpha) m_t(d\alpha)$ is right-continuous w.p.1. Since we are only concerned in this paper with the use of $\hat{A}^{m,\varepsilon} q$ in an integral—to get the martingale property (6.4)—the forms for $\hat{A}^{m,\varepsilon}$ given in the text are valid even without the right continuity, as shown below.

If $\int \bar{G}(x, \alpha) m_t(d\alpha)$ is not right-continuous, we can get the correct $\hat{A}^{m,\varepsilon} q$ that makes (6.4) an $\mathcal{F}_t^\varepsilon$-martingale by an approximation procedure. Simply let $m^\varepsilon(\cdot)$ be replaced by a piecewise constant right-continuous admissible control $m^{\varepsilon,\Delta}$, and replace $x^\varepsilon(\cdot)$ by the associated process $x^{\varepsilon,\Delta}(\cdot)$. Calculate $\hat{A}^{m^{\varepsilon,\Delta},\varepsilon} q(x^{\varepsilon,\Delta}(t), t)$ for the functions $q(\cdot, \cdot)$ used in the text, then take limits in $q(x^{\varepsilon,\Delta}(t), t) - \int_0^t \hat{A}^{m^{\varepsilon,\Delta},\varepsilon} q(x^\varepsilon(s), s)\, ds$. This will yield the forms used in the text, which are, in fact, just what we get by assuming the right continuity of $\int \bar{G}(x, \alpha) m_t(d\alpha)$.

REFERENCES

[1] H. J. KUSHNER AND W. RUNGGALDIER, *Nearly optimal state feedback controls for stochastic systems with wide band noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 298–315.

[2] M. COX AND I. KARATZAS, *Stationary control of Brownian motion in several dimensions*, Adv. Appl. Prob., 18 (1985), pp. 531–561.

[3] 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.

[4] W. H. FLEMING, *Generalized solutions in optimal stochastic control*, in Differential Games and Control Theory III, E. Roxin, P. T. Liu, and R. L. Sternberg, eds., Marcel Dekker, New York, 1977.

[5] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, 1984.

[6] E. PARDOUX AND D. TALAY, *Stability of linear differential systems with parametric excitation*, Nonlinear Stochastic Dynamic Engineering Systems, G. I. Schueller and F. Ziegler, eds., Proc. IUTAM Symposium, Innsbruck, 1987, Springer-Verlag, Berlin, New York, 1988.

[7] H. J. KUSHNER, *Necessary and sufficient conditions for optimality for the average cost per unit time problem with a diffusion model*, SIAM J. Control Optim., 16 (1977), pp. 330–346.

[8] S. ETHIER AND T. KURTZ, *Markov Processes: Characterization and Convergence*, McGraw-Hill, New York, 1986.

[9] H. J. KUSHNER, *Direct averaging and perturbed test function methods for weak convergence*, in Stochastic Optimization, V. Arkin, A. Shiryayev, and R. Wets, eds., Lecture Notes in Control and Information Science, 81, Springer-Verlag, Berlin, 1986.

[10] H. J. KUSHNER AND H. HUANG, *Approximate and limit results for non-linear filters with wide bandwidth observation noise*, Stochastics, 16 (1986), pp. 65–96.

[11] L. ARNOLD AND V. WIHSTUTZ, *Lyapunov exponents; a survey*, in Lyapunov Exponents, L. Arnold and V. Wihstutz, ed., Lecture Notes in Mathematics, 1186, Springer-Verlag, Berlin, New York, 1986.

[12] T. G. KURTZ, *Semigroups of conditioned shifts and approximations of Markov processes*, Ann. Probab., 4 (1975), pp. 618–642.

[13] G. BLANKENSHIP AND G. C. PAPANICOLAOU, *Stability and control of stochastic systems with wide bandwidth noise disturbances*, SIAM J. Appl. Math., 34 (1978), pp. 437–476.

[14] G. C. PAPANICOLAOU, D. STROOCK, AND S. R. S. VARADHAN, *Martingale approach to some limit theorems*, in Proc. 1976 Duke University Conference on Turbulence, Durham, NC, 1976.

[15] H. J. KUSHNER, *Jump diffusion approximations for ordinary differential equations with random right-hand sides*, SIAM J. Control Optim., 17 (1979), pp. 729–744.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 563-578, May 1989

DISTRIBUTED COMPUTATION OF NASH EQUILIBRIA IN LINEAR-QUADRATIC STOCHASTIC DIFFERENTIAL GAMES*

TAMER BAŞAR† AND SHU LI‡

Abstract. In this paper, a class of n-person stochastic linear-quadratic differential games under multiple probabilistic modeling is studied, with each player acquiring a noisy measurement of the initial state. Conditions for the existence and uniqueness of the Nash equilibrium is obtained, and a method is provided for an iterative distributed computation of the solution. The distributed algorithm involves learning in the policy space, and it does not require that each player knows the others' perceptions of the probabilistic model underlying the decision process. For the finite horizon problem, such an iteration converges whenever the length of the time horizon is sufficiently small, and the limit in this case is an affine policy for all players if the underlying distributions are jointly Gaussian. When the horizon is infinite and a discount factor is used in the cost functionals, the iteration converges under conditions depending on the magnitude of the discount factor, the limiting policies again being affine in the case of Gaussian distributions.

Key words. stochastic differential games, stable Nash equilibria, multiple probabilistic models, repeated incomplete games, distributed algorithms

AMS(MOS) subject classifications. 93E05, 90D25

1. Introduction.

The majority of the results in dynamic and differential game theory pertain to the so-called "complete games" where the players have a complete knowledge of the rules of the game and the cost functionals of other players, while "incomplete games" (cf. [1]-[3]) have attracted relatively little attention. One recent study on the latter class of problems is [4], which has developed a framework that allows different decision makers (synonymously, players) to adopt different (not necessarily consistent) probabilistic models. In this framework, [4] obtains conditions for existence, uniqueness, and stability of Nash and Stackelberg equilibria in quadratic static games and shows that under jointly Gaussian distributions, the Nash equilibrium is affine in the observations while the Stackelberg equilibrium policies are intrinsically nonlinear. Furthermore, it has been shown in [4] that multimodeling is well posed under the Nash solution concept (in the sense that the solution is "structurally continuous" in the limit as we go from multiple probabilistic models to a single model for all decision-makers), whereas it is structurally nonrobust under the Stackelberg solution concept.

In this paper, we extend the multiple probabilistic model of [4] to n-person stochastic differential games. The mathematical framework allows each player not to have access to the probability measure and parameters of the cost functionals adopted by other players, and the presence of a communication link between the players permits transmission of policy information, thereby enhancing learning in the policy space. We show that distributed computation and iteration in policy space leads to a unique Nash equilibrium when the finite time horizon is sufficiently short, and that this solution

* Received by the editors August 24, 1987; accepted for publication (in revised form) July 12, 1988. An earlier version of this paper was presented at the Tenth International Federation of Automatic Control (IFAC) World Congress on Automatic Control, Munich, Federal Republic of Germany, July 27-31, 1987.

† Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801. The research of this author was supported in part by Air Force Office of Scientific Research grant AFOSR 84-0056, and in part by Joint Services Electronics Program contract N00014-84-C-0149 through the University of Illinois, Urbana, Illinois 61801.

‡ Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138. Present address, Department of Systems and Industrial Engineering, University of Arizona, Tucson, Arizona 85721.

is affine for all players when the underlying distributions are jointly Gaussian. When the horizon is infinite and a discount factor is allowed, we show that the conditions depend on the magnitude of this discount factor.

In § 2 of the paper, we formulate the problem in precise mathematical terms in a Hilbert space setting, using a multiple probabilistic description; we also develop necessary and sufficient conditions for existence of an equilibrium solution. These conditions are further refined in § 3 in the proof of uniqueness and stability of the Nash solution and convergence of a related distributed algorithm. Section 3 also includes results on the special case when prior distributions are jointly Gaussian. Section 4 extends these results to the infinite horizon case, and the paper ends with the concluding remarks of § 5 and an Appendix.

2. Problem formulation and characterization of Nash equilibria. Let $(\Omega, \mathbf{F})$ be a measurable space, and let $\mathbb{P}$ denote the class of probability measures defined on it. Let $X = \mathbb{R}^m$ be the state space, let $Y_i = \mathbb{R}^{m_i}$ be the observation space for decision maker (DMi), $i \in \mathbb{N} := \{1, 2, \cdots, n\}$, and let $Z := X \times Y_1 \times \cdots \times Y_n$. Introduce $\mathbf{B}_z$ = Borel field of subsets of Z, $\mathbf{B}^k$ = Borel field of subsets of $\mathbb{R}^k$, $k = m, m_1, \cdots, m_n$, and random vectors $x_0 : (\Omega, \mathbf{F}) \to (X, \mathbf{B}^m)$, $y_i : (\Omega, \mathbf{F}) \to (Y_i, \mathbf{B}^{m_i})$, with corresponding Borel probability measures $\mathbf{P}_{x_0}$ and $\mathbf{P}_{y_i}$ induced by each $\mathbf{P} \in \mathbb{P}$.

Now, we consider a system modeled by an m-dimensional sample-path-continuous random process, satisfying the Itô stochastic differential equation

$$dx_t = \left(A(t)x_t + \sum_{i=1}^{n} B_i(t)u_{it} \right) dt + C(t)\, dw_t, \qquad t \geqq 0 \tag{1}$$

under the n probability measures $\mathbf{P}^1, \cdots, \mathbf{P}^n \in \mathbb{P}$, perceived by DM1, $\cdots$, DMn. Here x_0 is a random vector under the n probability measures $\mathbf{P}^1, \cdots, \mathbf{P}^n$, $x_0(\omega) \in X$; $\{w_t\}$ is an m-dimensional Brownian motion under $\mathbf{P}^1, \cdots, \mathbf{P}^n$; $A(t), B_1(t), \cdots, B_n(t), C(t)$ are appropriate dimensional matrices, continuous in $[0, t_f]$; $\{u_{it}\}$ is a p_i-dimensional random process denoting DMi's action, $i \in \mathbb{N}$; and $y_i(\omega) \in Y_i$ is the static measurement of DMi, $i \in \mathbb{N}$, related to the initial state x_0, with all these random quantities being well defined under the probability measures $\mathbf{P}^1, \cdots, \mathbf{P}^n$. Furthermore, $y_1, \cdots, y_n$ and x_0 are independent of $\{w_t\}$ under the n measures. Note that this is an open-loop information pattern, and one possible relationship between the y_is and the x_0 would be given by the linear measurement model:

$$y_i = H_i x_0 + v_i,$$

where H_i is an $m_i \times m$ dimensional matrix for each $i \in \mathbb{N}$, and $\{v_i, i \in \mathbb{N}\}$ is a sequence of independent random vectors, defined under each of the measures $\mathbf{P}^1, \cdots, \mathbf{P}^n$. We will have occasion to use this specific model later in § 3.

A permissible decision rule for DMi is a Borel-measurable mapping $\gamma_i : [0, t_f] \times R^{m_i} \to R^{p_i}$ such that

$$E^i \left\{ \int_0^{t_f} \| \gamma_i(t, y_i) \|^2 \, dt \right\} < \infty$$

where the norm $\|\cdot\|$ is taken as the Euclidean norm on R^{p_i}, for $i \in \mathbb{N}$, and E^i refers to unconditional expectation under the probability measure $\mathbf{P}^i$. Let Γ_i denote the set of all such permissible decision rules (strategies) for DMi, $i \in \mathbb{N}$.

Define the quadratic cost functional for DMi as follows:

$$J_i(\gamma) = E^i \left\{ \frac{1}{2} x_{t_f}^T Q_{if} x_{t_f} + \int_0^{t_f} \left(\frac{1}{2} x_t^T Q_i(t) x_t + \frac{1}{2} \sum_{j \in N} u_{jt}^T R_{ij}(t) u_{jt} \right) dt \,\middle|\, u_{lt} = \gamma_l(t, y_l), l \in \mathbb{N} \right\},$$

$$i \in \mathbb{N},$$

where $Q_{if} \geqq 0$, $Q_i(t) \geqq 0$ for all t, $R_{ii}(t) > 0$ for all t, for $i \in \mathbb{N}$, "T" denotes the transpose and $\gamma := (\gamma_1, \cdots, \gamma_n)$. Note that the above specifications make J_i strictly convex in γ_i for each $i \in \mathbb{N}$. Furthermore, for each J_i to be well defined and finite for all $(\gamma_1, \cdots, \gamma_n) \in \Gamma_1 \times, \cdots, \times \Gamma_n$, and under the given n probability measures, we have to make the following two basic assumptions, as in [4].

Assumption 2.1. For each fixed $i \in \mathbb{N}$, $\mathbf{P}^j_{y_i}, j \neq i, j \in \mathbb{N}$, *are absolutely continuous with respect to* $\mathbf{P}^i_{y_i}$.

Assumption 2.2. The Radon-Nikodym derivatives [5] satisfy, for some non-negative scalars h^i_{ji}

$$g^i_{ji}(\xi) = \mathbf{P}^j_{y_i}(d\xi)/\mathbf{P}^i_{y_i}(d\xi) \leqq h^i_{ji} < \infty$$

uniformly in ξ almost everywhere $\mathbf{P}^i_{y_i}$, for $j \neq i$, $i, j \in \mathbb{N}$.

To establish existence and uniqueness results, we also introduce the Hilbert spaces $\mathbf{X}_1, \cdots, \mathbf{X}_n$ and $\mathbf{\Gamma}_1, \cdots, \mathbf{\Gamma}_n$ as follows. First, we let $\mathbf{X}_i$ denote the completion of the space of continuous functions from $[0, t_f] \times \Omega$ into X, under the inner product

$$\langle x^1, x^2 \rangle_{\mathbf{x}_i} = E^i \left\{ \int_0^{t_f} x_t^{1^T} x_t^2 \, dt + x_{t_f}^{1^T} x_{t_f}^2 \right\}, \qquad i \in \mathbb{N}.$$

$\mathbf{\Gamma}_i$, on the other hand, is the policy space for DMi, defined as earlier but now also endowed with the inner product

$$\langle \gamma, \beta \rangle_{\mathbf{\Gamma}i} = E^i \left\{ \int_0^{t_f} \gamma^T(t, y_i) \beta(t, y_i) \, dt \right\}, \qquad i \in \mathbb{N}.$$

Finally we let $\mathbf{\Gamma} := \mathbf{\Gamma}_1 \times, \cdots, \times \mathbf{\Gamma}_n$ and $\gamma_{-i} := (\gamma_1, \cdots, \gamma_{i-1}, \gamma_{i+1}, \cdots, \gamma_n)$. We are now in a position to introduce in precise terms (in Definition 2.1 below) the equilibrium solution adopted in this paper, and also (in Definition 2.2) a refined version to be studied in § 3.

DEFINITION 2.1. An n-tuplet of policies $(\gamma_1^N, \cdots, \gamma_n^N) := \gamma^N \in \mathbf{\Gamma}$ is in Nash equilibrium if

$$J_i(\gamma^N) \leqq J_i(\gamma_i, \gamma^N_{-i}) \quad \forall \gamma_i \in \mathbf{\Gamma}_i, \quad i \in \mathbb{N}.$$

DEFINITION 2.2. A Nash equilibrium solution $\gamma^N \in \mathbf{\Gamma}$ is *stable* (with respect to a parallel update of the players' strategies) if for all $\gamma^{(0)} \in \Gamma$,

$$\gamma_i^N = \lim_{k \to \infty} \gamma_i^{(k)}, \qquad i \in \mathbb{N},$$

where

$$\gamma_i^{(k+1)} = \arg\min_{\gamma_i \in \Gamma_i} J_i(\gamma_i, \gamma_{-i}^{(k)}), \qquad i \in \mathbb{N}, \quad k = 0, 1, 2, \cdots.$$

Remark 2.1. It is important to note that the above is one possible definition of a stable Nash equilibrium, which is the one we are going to adopt in this paper. Here, the players update on their strategies in parallel, with each player using the most recently computed (and announced) strategies of the other players as fixed and given. There could be other definitions of stability, where the players update their strategies following a particular sequential order (see, for example, [6]). However, the different sequences generated by these different schemes will all have the same limit (provided that they converge), even though each may require a different condition of convergence, some being more restrictive than others. The parallel scheme adopted in this paper

seems to be well suited for an analysis in the framework of repeated games, and also it enables us to obtain rather strong results on the existence of a unique Nash equilibrium. It is needless to say that, since the definition of a stable Nash equilibrium requires convergence for all starting strategies $\gamma_i^{(0)}$, $i \in \Gamma$, a stable Nash equilibrium is necessarily *unique*.

Remark 2.2. The formulation above models a decision scenario where the players have different (and necessarily inconsistent) perceptions of the statistics of the underlying stochastic phenomena, and stay with this perception until the end of the decision process. However, to completely characterize (and compute) the Nash equilibrium solution, all this statistical information and the cost structure of each player have to be common information to all. In view of this, what the algorithm presented in Definition 2.2 accomplishes is to free the players from the task of acquiring all this information, at the expense of repeating the game a number of times and by transmitting (only) policy information from one stage to another. At each stage, DMi uses the same measurement y_i he acquired at stage zero, and in this sense there is no learning with respect to the random variables involved, nor is there any learning with respect to the measurements made by the other players (i.e., there is no learning in action space); however, if this had been the case, then the convergent solution of the sequence of repeated games would have no relationship with the solution of the original stochastic differential game. In our formulation, learning takes place only in the policy space, ensuring the players to arrive at the (Nash) equilibrium (in the limit), with this end result depending explicitly on the different probabilistic perceptions and the cost functions of *all* players (even though this information is never shared by them).

We will now first focus on the derivation of necessary and sufficient conditions for Nash equilibria, without giving consideration to stability. This is done below in Theorem 2.1. Subsequently, in § 3, we turn to the stability analysis of the solution, using the iterative scheme of Definition 2.2.

THEOREM 2.1. *For the continuous-time decision problem formulated above, there exists a Nash equilibrium solution if and only if there exist* $2n$ *processes* $\hat{x}_{it}$, λ_{it}, $i \in \mathbb{N}$, *satisfying the following set of coupled differential equations with mixed boundary conditions*:

$$\begin{pmatrix} \dot{\hat{x}}_{it} \\ \dot{\lambda}_{it} \end{pmatrix} = \begin{pmatrix} A & -B_i R_{ii}^{-1} B_i^T \\ -Q_i & -A^T \end{pmatrix} \begin{pmatrix} \hat{x}_{it} \\ \lambda_{it} \end{pmatrix} + \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} \begin{pmatrix} B_j \\ 0 \end{pmatrix} \hat{u}_{jit}, \tag{2}$$

$$\hat{x}_{i0} = E^i(x_0 | y_i), \qquad \lambda_{it_f} = Q_{if} \hat{x}_{it_f},$$

where

$$\hat{u}_{jit} := E^i\{\gamma_j^N(t, y_j) | y_i\}, \qquad j \neq i, \quad i, j \in \mathbb{N}.$$

If (2) *admits a solution subject to the given boundary conditions, then the set of policies*

$$\gamma_i^N(t, y_i) = -R_{ii}^{-1} B_i^T \lambda_{it}(y_i) \quad a.e.\ \mathbf{P}^i, \quad i \in \mathbb{N} \tag{3}$$

constitutes a Nash equilibrium solution.

Proof. We prove the theorem by using a variational approach. For fixed i, and fixed $\gamma, \beta \in \mathbf{\Gamma}$, perturb $\gamma_i \to \gamma_i + \varepsilon\beta_i$ where ε is a nonzero scalar; then, correspondingly, $x_t \to x_t + \varepsilon\eta_{it}$, $\eta_i \in \mathbf{X}_i$. Using this in (1), we obtain

$$\dot{\eta}_{it} = A\eta_{it} + B_i\beta_i(t, y_i), \quad \eta_{i0} = 0, \quad i \in \mathbb{N}. \tag{4}$$

Now,

$$\begin{aligned}\Delta J_i &= J_i(\gamma_i + \varepsilon\beta_i, \gamma_{-i}) - J_i(\gamma_i, \gamma_{-i}) \\ &= E^i\left\{ x_{t_f}^T Q_{if}\eta_{it_f} + \int_0^{t_f} (x_t^T Q_i\eta_{it} + u_{it}^T R_{ii}\beta_i(t, y_i))\, dt \right\}\varepsilon \\ &\quad + \frac{1}{2} E^i\left\{ \eta_{it_f}^T Q_{if}\eta_{it_f} + \int_0^{t_f} (\eta_{it}^T Q_i\eta_{it} + \beta_i^T(t, y_i) R_{ii}\beta_i(t, y_i))\, dt \right\}\varepsilon^2.\end{aligned}$$

Since $Q_{if} \geqq 0$, $Q_i(\cdot) \geqq 0$, $R_{ii}(\cdot) > 0$, the second term above is positive for all $\beta_i \in \mathbf{\Gamma}_i$, $\beta_i \neq 0$, and hence a necessary and sufficient condition for (γ_i, γ_{-i}) to constitute a Nash equilibrium solution is that

$$(5)\quad \forall \beta_i \in \mathbf{\Gamma}_i, \quad \delta J_i = E^i\left\{ E^i(x_{t_f}^T | y_i) Q_{if}\eta_{it_f} + \int_0^{t_f} [E^i(x_t^T | y_i) Q_i\eta_{it} + u_{it}^T R_{ii}\beta_i(t, y_i)]\, dt \right\} = 0, \quad i \in \mathbb{N}.$$

Now, by (4),

$$(6)\qquad \eta_{it} = \int_0^t \Phi(t, \tau) B_i\beta_i(\tau, y_i)\, d\tau =: T_{Ai}\beta_i, \qquad i \in \mathbb{N},$$

where $T_{Ai} : \mathbf{\Gamma}_i \to \tilde{\mathbf{X}}_i \subset \mathbf{X}_i$; $\tilde{\mathbf{X}}_i = \{x \in \mathbf{X}_i, x \text{ is a } y_i\text{-measurable (under } \mathbf{P}^i\text{) stochastic process}\}$, and $\dot{\Phi}(t, s) = A(t)\Phi(t, s)$, $\Phi(s, s) = I$, i.e., Φ is the state transition matrix associated with A. It is not hard to see that the adjoint of T_{Ai} is

$$(7)\qquad (T_{Ai}^* x)_t = B_i^T(t) \int_t^{t_f} \Phi^T(\tau, t) x_\tau\, d\tau + B_i^T(t)\Phi^T(t_f, t) x_{t_f}.$$

First substituting (6) into (5), and then using (7) in the resulting expression under the inner product defined earlier on $\mathbf{X}_i$, we obtain for all $\beta_i \in \mathbf{\Gamma}_i$ (cf. [7]):

$$E^i\left\{ \left\langle B_i^T(t) \int_t^{t_f} \Phi^T(\tau, t) Q_i(\tau) E^i(x_\tau | y_i)\, d\tau + B_i^T(t)\Phi^T(t_f, t) Q_{if} E^i(x_{t_f} | y_i) + R_{ii}u_{it}, \beta_i \right\rangle_{\mathbf{\Gamma}_i} \right\} = 0$$

which holds if and only if

$$u_{it} = -R_{ii}^{-1} B_i^T \lambda_{it}, \quad \text{a.e. } \mathbf{P}^i$$

where

$$\lambda_{it} := \int_t^{t_f} \Phi^T(\tau, t) Q_i(\tau) E^i(x_\tau | y_i)\, d\tau + \Phi^T(t_f, t) Q_{if} E(x_{t_f} | y_i).$$

The latter can equivalently be written as

$$\dot{\lambda}_{it} = -Q_i\hat{x}_{it} - A^T\lambda_{it}, \qquad \lambda_{it_f} = Q_{if}\hat{x}_{it_f},$$

where $\hat{x}_{it}$ is the conditional mean of x_t under the measure $\mathbf{P}_i$, given y_i and the set of fixed policies $\gamma \in \mathbf{\Gamma}$:

$$\hat{x}_{it} := E^i(x_t | y_i).$$

For each $i \in \mathbb{N}$ it satisfies the differential equation:

$$\dot{\hat{x}}_{it} = A\hat{x}_{it} - B_i R_{ii}^{-1} B_i^T \lambda_{it} + \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} B_j \hat{u}_{jit}, \qquad \hat{x}_{i0} = E^i[x_0 | y_i],$$

with

$$\hat{u}_{jit} := E^i\{\gamma_j(t, y_j)\,|\,y_i\}, \qquad j \neq i, \quad i, j \in \mathbb{N}.$$

This completes the proof of the theorem. □

We observe that the condition given in the theorem is reminiscent of the one available for deterministic linear-quadratic differential games (cf. [6]), but here we have n different state equations, because of different perceptions of the players, and likewise n different co-state equations (see Remark 3.4 in the next section for further elaboration on this point).

3. A distributed algorithm and existence of a stable equilibrium. We now study further the condition of Theorem 2.1 and in particular the stability of the equilibrium solution, as in Definition 2.2. Recall that the iterative (distributed) computation of the Nash solution involves the sequences $\{\gamma_i^{(k)};\ k = 0, 1, \cdots\}$, $i \in \mathbb{N}$, generated by

$$\gamma_i^{(k+1)} = \arg\min_{\gamma_i \in \Gamma_i} J_i(\gamma_i, \gamma_{-1}^{(k)}), \qquad k = 0, 1, \cdots, \quad i \in \mathbb{N} \tag{8}$$

for fixed (but arbitrarily chosen) $\gamma^{(0)} \in \mathbf{\Gamma}$. Now, utilizing the main idea of the proof of Theorem 2.1, we can rewrite algorithm (8) equivalently as follows:

$$\gamma_i^{(k+1)}(t, y_i) = -R_{ii}^{-1} B_i^T \lambda_{it}^{(k)}(y_i), \qquad i \in \mathbb{N}, \tag{9}$$

$$\begin{pmatrix} \dot{\hat{x}}_{it}^{(k)} \\ \dot{\lambda}_{it}^{(k)} \end{pmatrix} = \begin{pmatrix} A & -B_i R_{ii}^{-1} B_i^T \\ -Q_i & -A^T \end{pmatrix} \begin{pmatrix} \hat{x}_{it}^{(k)} \\ \lambda_{it}^{(k)} \end{pmatrix} + \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} \begin{pmatrix} B_j \\ 0 \end{pmatrix} \hat{u}_{jit}^{(k)},$$

$$\begin{pmatrix} \hat{x}_{i0}^{(k)} \\ \lambda_{it_f}^{(k)} \end{pmatrix} = \begin{pmatrix} E^i(x_0\,|\,y_i) \\ Q_{if} \hat{x}_{it_f}^{(k)} \end{pmatrix}, \qquad i \in \mathbb{N} \tag{10}$$

where $\hat{u}_{jit}^{(k)} = E^i\{\gamma_j^{(k)}(t, y_j)\,|\,y_i\}$. Hence, to implement the learning scheme, at each stage $k+1$, DMi chooses $\gamma_i^{(k+1)}$ as the optimal response to $\gamma_j^{(k)}$, $j \in \mathbb{N}$, $j \neq i$, by (9) and (10). Our goal is to obtain conditions under which the sequences $\{\gamma_i^{(k)}\}$, $i \in \mathbb{N}$ converge, with the limit necessarily being the n-tuplet $(\gamma_i^N,\ i \in \mathbb{N})$, which constitutes the unique Nash solution to the original game. Note that if such conditions can be found, then the players can achieve the *stable* Nash solution through the given learning process (cf. Remark 2.2), even though DMi does not know $\mathbf{R}_{jj}$, Q_j, and $\mathbf{P}^j$, $j \neq i$, $i, j \in \mathbb{N}$.

First, by the transition matrix method, we can solve (2) and obtain the solution to be the following lemma.

LEMMA 3.1. *The solution to* (2) *is given by*

$$\begin{aligned} \hat{x}_{it} &= X_i(t) X_i^{-1}(0) \hat{x}_{i0} + \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} \int_0^t F_i(t, \tau) B_j(\tau) \hat{u}_{ji\tau}\, d\tau, \\ \lambda_{it} &= \Lambda_i(t) X_i^{-1}(0) \hat{x}_{i0} + \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} \int_t^{t_f} G_i(t, \tau) B_j(\tau) \hat{u}_{ji\tau}\, d\tau \end{aligned} \tag{11}$$

where

$$\begin{bmatrix} X_i(t) \\ \Lambda_i(t) \end{bmatrix} = \begin{bmatrix} \hat{x}_{it}^{(1)} \cdots \hat{x}_{it}^{(m)} \\ \lambda_{it}^{(1)} \cdots \lambda_{it}^{(m)} \end{bmatrix};$$

and for each s, $(\hat{x}_{it}^{(s)}, \lambda_{it}^{(s)})$ *is a solution of* (2) *with zero input* (*i.e.*, $\hat{u}_{jit} \equiv 0$), *and with the initial condition*

$$\hat{x}_{i0}^{(s)} = [0, \cdots, 0, 1, 0, \cdots]^T$$

where the s*th component is one*; *and* $\lambda_{if}^{(s)} = Q_{if} x_{if}^{(s)}$.

Furthermore, $F_i(t,\tau)$, $G_i(t,\tau)$ satisfy the following equations:

(i)
$$\frac{dF_i(t,\tau)}{dt} = A(t)F_i(t,\tau) - B_i(t)R_{ii}^{-1}(t)B_i^T(t)G_i(t,\tau), \qquad F_i(t,t) = I,$$
$$\frac{dG_i(t,\tau)}{dt} = -Q_i(t)F_i(t,\tau) - A^T(t)G_i(t,\tau), \qquad G_i(t,t) = 0;$$

(ii)
$$\begin{cases} F_i(t,\tau) = 0 & \text{for } t < \tau \leqq t_f, \\ G_i(t,\tau) = 0 & \text{for } 0 \leqq \tau \leqq t. \end{cases}$$

Note that in (11), $\hat{x}_{it}$ *denotes the conditional mean of the state vector* x_t, *as perceived by* DM*i*.

Proof. See the Appendix for the proof of Lemma 3.1. □

Next, we substitute (11) into (9) to obtain, for each $i \in \mathbb{N}$,

$$\gamma_i^{(k+1)}(t, y_i) = \xi_i(t, y_i) + (\Psi_i \gamma^{(k)})_t \tag{12}$$

where

$$\xi_i(t, y_i) = -R_{ii}^{-1}(t)B_i^T(t)\Lambda_i(t)X_i^{-1}(0)\hat{x}_{i0}, \qquad \hat{x}_{i0} = E^i(x_0 | y_i), \tag{13a}$$

and $\Psi_i : \mathbf{\Gamma} \to \mathbf{\Gamma}_i$ is the linear operator

$$(\Psi_i \gamma)_t = \int_t^{t_f} K_i(t,\tau)\hat{u}^i(\tau)\, d\tau \tag{13b}$$

with

$$K_i(t,\tau) := -R_{ii}^{-1}(t)B_i^T(t)G_i(t,\tau), \tag{14a}$$

$$\hat{u}^i(\tau) := \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} B_j(\tau)E^i[\gamma_j(\tau, y_j) | y_i]. \tag{14b}$$

Let $\Psi : \Gamma \to \Gamma$ be defined by

$$\Psi = [\Psi_1^T, \cdots, \Psi_n^T]^T.$$

Then, the convergence of the distributed algorithm (8) (or, equivalently, of the sequence generated by (12)) is equivalent to existence of a limit, in the Hilbert space $\mathbf{\Gamma}$, to the sequence $\{\gamma^{(k)}\}$ generated by

$$\gamma^{(k+1)} = \Psi\gamma^{(k)}, \qquad k = 0, 1, \cdots, \tag{15}$$

for arbitrary $\gamma^{(0)} \in \mathbf{\Gamma}$. Such a limit is guaranteed to exist if, for some $\alpha \in [0, 1)$,

$$\|\Psi_\gamma\|_{\mathbf{\Gamma}}^2 = \sum_{i \in \mathbb{N}} \|\Psi_i\gamma\|_{\mathbf{\Gamma}_i}^2 \leqq \alpha^2 \|\gamma\|_{\mathbf{\Gamma}}^2 \equiv \alpha^2 \sum_{i \in \mathbb{N}} \|\gamma_i\|_{\mathbf{\Gamma}_i}^2, \tag{16}$$

this result being a direct consequence of the contraction mapping theorem in complete metric spaces (cf. [7]). Our main result below makes this condition precise.

THEOREM 3.1. *Let functions* $a_i(s)$ *and* $\lambda^i{}_B(s)$, $i \in \mathbb{N}$, *be defined on* $[0, t_f]$ *as follows*:

$$a_i^2(t) := \int_t^{t_f} \operatorname{tr}[K_i(t,s)K_i^T(t,s)]\, ds, \qquad t \in [0, t_f], \tag{17a}$$

$$\lambda^i{}_B(s) := \lambda_{\max}[b^{iT}(s)b^i(s)] \tag{17b}$$

where

$$b^i(s) := [B_1(s), \cdots, B_{i-1}(s), 0, B_{i+1}(s), \cdots, B_n(s)], \tag{17c}$$

$\lambda_{\max}[\cdot]$ *denotes the largest eigenvalue of the matrix* $[\cdot]$, *and* $K_i(t, s)$ *is as defined by* (14a). *Let there exist a scalar* α, *such that*

$$\alpha^2 := \max_{\substack{i,s \\ i\in N \\ s\in[0,t_f]}} \sum_{\substack{j\in\mathbb{N} \\ j\neq i}} h_{ji}^i \lambda_B^i(s) \int_0^s a_j^2(t)\, dt < 1, \tag{18}$$

where h_{ji}^i *are as introduced in Assumption* 2.2.

Then,

(i) *There exists a unique Nash equilibrium solution*;

(ii) *The solution is stable, i.e., the distributed algorithm* (8) (*equivalently* (12)) *converges to the Nash equilibrium regardless of the choice of initial strategies.*

Proof. Since Ψ_i in (12) is a linear operator, it will be sufficient to show that under the hypotheses of the theorem, Ψ is a contraction mapping [7]. Toward this end, first consider the quantity

$$(*) \qquad \|\Psi_i\gamma\|_{\Gamma_i}^2 = \int_0^{t_f} dt\, E^i\left\{\left\|\int_t^{t_f} K_i(t,\tau)\hat{u}^i(\tau)\, d\tau\right\|^2\right\}$$

where as before, $\|\cdot\|$ (without a subscript) denotes the Euclidean norm. To show that this quantity is bounded above by

$$(**) \qquad \int_0^{t_f} dt\, a_i^2(t) \int_t^{t_f} E^i\{\|\hat{u}^i(\tau)\|^2\}\, d\tau.$$

we introduce the partitioned representation

$$K_i(t, s) = [k_{i1}^T(t, s), \cdots, k_{ip_i}^T(t, s)]^T$$

and note that the integrand of (*) can be written as follows:

$$E^i\left(\sum_{j=1}^{p_i}\left[\int_t^{t_f} d\tau\, k_{ij}(t,\tau)\hat{u}^i(\tau)\right]^2\right\} \leqq E^i\left\{\sum_{j=1}^{p_i}\left(\int_t^{t_f} d\tau\|k_{ij}(t,\tau)\|^2\right)\int_t^{t_f} d\tau\|\hat{u}^i(\tau)\|^2\right\}$$

$$= \int_t^{t_f}\sum_{j=1}^{p_i}\|k_{ij}(t,s)\|^2\, ds \cdot \int_t^{t_f} E^i\{\|\hat{u}^i(\tau)\|^2\}\, d\tau,$$

which is equivalent to the integrand of (**) since the first (product) term above is $a_i^2(t)$ by (17a). We should note that in arriving at the inequality above we have used the Cauchy–Schwarz inequality (cf. [7]) (twice) in two different inner product spaces. Next, let us concentrate on the inner integrand of (**):

$$E^i\{\|\hat{u}^i(s)\|^2\} = E^i\left\{\left\|\sum_{\substack{j\in\mathbb{N} \\ j\neq i}} B_j(s)E^i[\gamma_j(s, y_j)\,|\,y_i]\right\|^2\right\}$$

(majorizing by the largest eigenvalue of the weighting matrix)

$$\leqq \lambda_B^i(s)E^i\left\{\sum_{\substack{j\in\mathbb{N} \\ j\neq i}}\|E^i[\gamma_j(s, y_j)\,|\,y_i]\|^2\right\}$$

(nonexpansive property of conditional expectation (cf. [5])

$$\leqq \lambda_B^i(s)E^i\left\{\sum_{\substack{j\in\mathbb{N} \\ j\neq i}}\|\gamma_j(s, y_j)\|^2\right\}$$

(change of measure)

$$= \lambda^i_B(s) \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} E^j\{g^j_{ij}(y_j)\|\gamma_j(s,y_j)\|^2\} \leqq \lambda^i_B \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^j_{ij} E^j\{\|\gamma_j(s,y_j)\|^2\}.$$

Using this bound in (**) we can thus bound (*) by

$$\|\Psi_i\gamma\|^2_{\Gamma_i} \leqq \int_0^{t_f} dt\, a_i^2(t) \int_t^{t_f} ds\, \lambda^i_B(s) \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^i_{ij} E^j\{\|\gamma_j(s,y_j)\|^2\}$$

$$= \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^j_{ij} \int_0^{t_f} ds\, \lambda^i_B(s) E^j\{\|\gamma_j(s,y_j)\|^2\} \int_0^s a_i^2(t)\, dt.$$

Now, finally,

$$\|\Psi\gamma\|^2_{\Gamma} = \sum_{i=1}^n \|\Psi_i\gamma\|^2_{\Gamma_i} \leqq \sum_{i=1}^n \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^j_{ij} \int_0^{t_f} ds\, \lambda^i_B(s) E^j\{\|\gamma_j(s,y_j)\|^2\} \int_0^s a_i^2(t)\, dt$$

$$= \int_0^{t_f} \sum_{i=1}^n E^i\{\|\gamma_i(s,y_i)\|^2\}\, ds \cdot \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^i_{ji} \lambda^j_B(s) \int_0^s a_j^2(t)\, dt$$

$$\leqq \int_0^{t_f} \sum_{i=1}^n E^i\{\|\gamma_i(s,y_i)\|^2\}\, ds \cdot \alpha^2 = \alpha^2 \|\gamma\|^2_{\Gamma}$$

where, in arriving at the next to the last line, we have interchanged the order of the two summations, and in the last we have used (18). This, then, shows that Ψ is a contraction mapping, thus completing the proof of the theorem (using the Banach contraction mapping theorem (cf. [7]). □

Remark 3.1. A courser bound for (18) is

$$\alpha^2 \leqq t_f \max_{i,s} \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h^i_{ji} \lambda^j_B(s) \bar{a}_j^2$$

where $\bar{a}_j^2 := \max_s a_j^2(s)$. This shows that there always exists a t_f, sufficiently small, such that the learning process converges.

Remark 3.2. In general, it is difficult to solve the two-point boundary value problem (2) explicitly (other than to obtain the implicit representation provided in Lemma 3.1) because we cannot "close the loop" by substituting $\gamma_i(t, y_i) = -R_{ii}^{-1} B_i^T \lambda_{it}(y_i)$ into the differential equations. The presence of the conditional expectation term $E^i\{\lambda_{jt}|y_i\}$ (which is the stochastic product term of $\hat{u}_{jit}\}$ contributes to this difficulty. When the distributions are jointly Gaussian, however, certain simplifications are possible, and there emerges the possibility of obtaining the solution analytically, as we discuss next.

Now, let us consider the special class of problems in which the joint distributions of $x_0, y_1, \cdots, y_n$ are zero-mean Gaussian under all n measures. In this case there exist matrices Σ_{0i}, Σ_{ij}, of appropriate dimensions, such that[1]

$$\hat{x}_{i0} = E^i[x_0|y_i] = \Sigma_{0i} y_i, \qquad E^i[\hat{x}_{j0}|y_i] = \Sigma_{ji}\hat{x}_{i0}, \qquad i \neq j, \quad i,j \in \mathbb{N}.$$

[1] This would be true if, for example, $y_i = H_i x_0 + v_i$, $i \in N$, as given in § 2, where $x_0, v_1, \cdots, v_n$ are independent Gaussian random vectors under the measures $\mathbf{P}^1, \cdots, \mathbf{P}^n$.

We introduce matrix functions $M_i(m \times m)$ and $N_i(m \times m)$ on $[0, t_f]$, $i \in N$, satisfying the differential equations

$$(19a) \qquad [\dot{M}_i + M_i A + A^T M_i + Q_i - M_i B_i R_{ii}^{-1} B_i^T M_i] N_i - \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} M_i B_j R_{jj}^{-1} B_j^T M_j N_j \Sigma_{ji} = 0,$$

$$M_i(t_f) = Q_{i_f},$$

$$(19b) \qquad \dot{N}_i = A N_i - B_i R_{ii}^{-1} B_i^T M_i N_i - \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} B_j R_{jj}^{-1} B_j^T M_j N_j \Sigma_{ji}, \qquad N_i(0) = I.$$

Then the following result follows as a corollary to Theorem 3.1.

COROLLARY 3.1. *When the underlying distributions are zero-mean Gaussian under the n measures we have the following*:

(i) *There exists a unique stable Nash equilibrium solution, linear in the measurements, provided that* $\alpha^2 < 1$;

(ii) *The unique linear solution is given by*

$$(20) \qquad \gamma_i^N(t, y_i) = -R_{ii}^{-1}(t) B_i^T(t) M_i(t) \hat{x}_{it} \quad a.e.\ \mathbf{P}^i, \quad i \in \mathbb{N},$$

where

$$(21) \qquad \hat{x}_{it} = N_i(t) \Sigma_{0i} y_i, \qquad i \in \mathbb{N},$$

provided that (19) *admits a solution.*

Proof. Linearity of the solution alluded to in Theorem 3.1 follows readily from recursions (9)-(10) by taking $\gamma_i^{(0)}(y_i)$ to be any linear function of y_i, $i \in \mathbb{N}$, and seeing that, since the underlying distributions are jointly Gaussian under all n measures, $\hat{x}_{it}^{(k)}$ and $\lambda_{it}^{(k)}$ are linear in y_i for all $i \in \mathbb{N}$, $k = 0, 1, \cdots$. For a proof of part (ii) of the corollary, it is sufficient to note that if (19) admits a solution, then $\hat{x}_{it} = N_i(t) \Sigma_{0i} y_i$ and $\lambda_{it} = M_i(t) \hat{x}_{it}$, $i \in \mathbb{N}$, solve (2). □

Remark 3.3. Even though (19) is a complicated set of differential equations, it does offer the advantage of computing the unique Nash policies off-line, as opposed to the general result of Lemma 3.1 which requires on-line computation (for the non-Gaussian case).

Remark 3.4. For the special case when the players make identical measurements, $y_1 = \cdots = y_n \equiv y$, we have $\Sigma_{ji} = I$ (but still Σ_{0i} is dependent of the index i, since the subjective probabilities are different), under which (19a) and (19b) simplify, with $N_i(t) = N(t)$, to

$$(19a') \qquad \dot{M}_i + M_i A + A^T M_i + Q_i - M_i B_i R_{ii}^{-1} B_i^T M_i - M_i \sum_{\substack{j \in \mathbb{N} \\ j \neq i}} B_j R_{jj}^{-1} B_j^T M_j = 0,$$

$$M_i(t_f) = Q_{if},$$

$$(19b') \qquad \dot{N} = \left(A - \sum_{j \in \mathbb{N}} B_j R_{jj}^{-1} B_j^T M_j \right) N, \qquad N(0) = I,$$

where in arriving at (19a') from (19a) we have made use of the fact that the unique solution of (19b') for each fixed $\{M_j\}$ is a full-rank matrix function (in fact a state transition matrix). The above are precisely the matrix differential equations arising in the open-loop Nash solution of deterministic differential games (see, e.g., [6, Thm. A-2]); (19a') is the open-loop (coupled) matrix Riccati differential equation, and (19b') is the linear differential equation satisfied by the state transition matrix N.

4. Infinite horizon games. In this section we obtain counterparts of the results in § 3 for similarly structured infinite horizon stochastic differential games with discounted costs. We can cite at least two main reasons for studying such a problem. First, for differential games with long (but not infinite) horizon (be they repeated or not), the results presented here could provide a good approximation for the true Nash solution. Second, for genuine infinite-horizon stochastic differential games with different subjective probabilities for the players, the distributed learning algorithm presented here could provide an effective means of computing the Nash strategies.

With this prelude, we now go into a precise formulation of the problem. The system is still characterized by the stochastic differential equation (1), with the apparent modification that every quantity is defined on the interval $[0, \infty)$. The cost functions are

$$J_i(\gamma) = E^i\left\{\int_0^\infty e^{-\beta_i t}\left(\frac{1}{2}x_t^T Q_i(t)x_t + \frac{1}{2}\sum_{j\in\mathbb{N}} u_{jt}^T R_{ij}(t)u_{jt}\right)dt \,\middle|\, u_{lt} = \gamma_l(t, y_l),\ l\in\mathbb{N}\right\},$$

$$i\in\mathbb{N},$$

where $\beta_i > 0$, $i\in\mathbb{N}$ are the discount factors. A permissible policy for *DMi* is a Borel-measurable mapping $\gamma_i : [0, \infty)\times\mathbb{R}^{m_i}\to\mathbb{R}^{P_i}$ such that

$$E^i\left\{\int_0^\infty e^{-\beta_i t}\|\gamma_i(t, y_i)\|^2\,dt\right\} < \infty.$$

The policy space $\mathbf{\Gamma}_i$ for *DMi* is now the Hilbert space of all permissible policies, under the inner product

$$\langle\gamma, \beta\rangle_{\mathbf{\Gamma}_i} := E^i\left\{\int_0^\infty e^{-\beta_i t}\gamma^T(t, y_i)\beta(t, y_i)\,dt\right\}, \qquad i\in\mathbb{N}.$$

Similarly, we define the Hilbert spaces $\mathbf{X}_i$, $i\in\mathbb{N}$, as the completion of the space of continuous functions from $[0, \infty)\times\Omega$ into $\mathbf{X}$, under the inner products

$$\langle x^1, x^2\rangle_{x_i} := E^i\left\{\int_0^\infty e^{-\beta_i t}[x_t^{1T}x_t^2]\,dt\right\}.$$

Clearly, in order that the state trajectory $\{x_t\}$ be in $\mathbf{X}_i$, we need the following assumption:

Assumption 4.1. The system parameters $A, B_1, \cdots, B_n, C$, and the discount factors β_i are such that for all $\gamma_i\in\mathbf{\Gamma}_i$, the resulting trajectory $\{x_t\}$ (that is, the solution of (1)) satisfies the boundedness condition

$$E^i\left\{\int_0^\infty e^{-\beta_i t}x_t^T x_t\,dt\right\} < \infty \quad \text{for } i\in\mathbb{N}.$$

Now we state the following two results, which are the counterparts of Theorem 2.1 and Lemma 3.1, respectively.

THEOREM 4.1. *Suppose that Assumption* 4.1 *is satisfied, and that*

$$\lim_{t\to+\infty}\int_t^\infty e^{-\beta_i(\tau-t)}\Phi^T(\tau, t)Q_i(\tau)E^i(x_{i\tau}\,|\,y_i)\,dt = 0, \qquad i\in\mathbb{N}, \tag{2.3}$$

where $\Phi(\tau, t)$ is the state transition matrix function associated with A.

Then, the infinite-horizon stochastic differential game admits a Nash equilibrium solution if, and only if, there exist $2n$ processes $\hat{x}_{it}$, λ_{it}, $i\in\mathbb{N}$, satisfying the following set

of coupled differential equations with mixed boundary conditions:

$$\begin{bmatrix}\dot{\hat{x}}_{it}\\ \dot{\lambda}_{it}\end{bmatrix}=\begin{bmatrix}A & -B_iR_{ii}^{-1}B_i^T\\ -Q_i & \beta_iI-A^T\end{bmatrix}\begin{bmatrix}\hat{x}_{it}\\ \lambda_{it}\end{bmatrix}+\sum_{\substack{j\in\mathbb{N}\\ j\neq i}}\begin{bmatrix}B_j\\ 0\end{bmatrix}\hat{u}_{jit},$$

$$\hat{x}_{i0}=E^i(x_0|y_i),\quad \lim_{t\to\infty}\lambda_{it}=0,\quad i\in\mathbb{N}, \tag{24}$$

where

$$\hat{u}_{jit}:=E^i\{\gamma_j^N(t,y_j)|y_i\},\qquad j\neq i,\quad i,j\in\mathbb{N}.$$

If (24) *admits a solution subject to the given boundary conditions, then the set of policies*

$$\gamma_i^N(t,y_i)=R_{ii}^{-1}(t)B_i^T(t)\lambda_{it}(y_i)\quad a.e.\ \mathbf{P}^i,\quad i\in\mathbb{N},$$

constitutes a Nash equilibrium solution.

LEMMA 4.1. *The solution to* (24) *is given by*

$$\hat{x}_{it}=\hat{x}_{it}^{(h)}+\sum_{\substack{j\in\mathbb{N}\\ j\neq i}}\int_0^t\bar{F}_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau, \tag{25a}$$

$$\lambda_{it}=\lambda_{it}^{(h)}+\sum_{\substack{j\in\mathbb{N}\\ j\neq i}}^{\infty}\int_t^{\infty}\bar{G}_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau \tag{25b}$$

where $(\hat{x}_{it}^{(h)},\lambda_{it}^{(h)})$ *is a solution to the homogeneous part (assuming* $\hat{u}_{jit}\equiv 0$*), and* $\bar{F}_i(t,\tau)$, $\bar{G}_i(t,\tau)$ *satisfy the following*:

(i)
$$\frac{d\bar{F}_i(t,\tau)}{dt}=A(t)\bar{F}_i(t,\tau)-B_i(t)R_{ii}^{-1}(t)B_i(t)\bar{G}_i(t,\tau),\qquad \bar{F}_i(t,t)=I,$$
$$\frac{d\bar{G}_i(t,\tau)}{dt}=-Q_i(t)\bar{F}_i(t,\tau)+(\beta_iI-A^T(t))\bar{G}_i(t,\tau),\qquad \bar{G}_i(t,t)=0;$$

(ii)
$$\bar{F}_i(t,\tau)=0\quad \textit{for } t<\tau<\infty,$$
$$\bar{G}_i(t,\tau)=0\quad \textit{for } 0<\tau\leqq t.$$

Now, the distributed algorithm replacing (12) in this case is

$$\gamma_i^{(k+1)}(t,y_i)=\bar{\xi}_i(t,y_i)+(\bar{\Psi}_i\gamma^{(k)})_t\qquad i\in\mathbb{N}, \tag{26}$$

where all terms are as defined before (without overbar), with the obvious change that $t_f\to\infty$. We therefore seek conditions for existence of a limit to the sequence $\{\gamma^{(k)}\}$ generated by the counterpart of (15):

$$\gamma^{(k+1)}=\bar{\Psi}\gamma^{(k)},\qquad k=0,1,\cdots$$

for arbitrary $\gamma^{(0)}\in\mathbf{\Gamma}$. Toward establishing the validity of (16) for some $\alpha^2<1$, we first proceed as in the proof of Theorem 3.1 and obtain the following sequence of equalities and inequalities:

$$\begin{aligned}\|\bar{\Psi}_i\gamma\|_{\mathbf{\Gamma}_i}^2&=\int_0^{\infty}e^{-\beta_it}E^i\left\{\left\|\int_t^{\infty}\bar{K}_i(t,\tau)\hat{u}^i(\tau)\,d\tau\right\|^2\right\}\\
&=\int_0^{\infty}e^{-\beta_it}E^i\left\{\sum_{j=1}^{p_2}\left[\int_t^{\infty}d\tau\,\bar{k}_{ij}(t,\tau)\hat{u}^i(\tau)\right]^2\right\}\\
&\leqq\int_0^{\infty}e^{-\beta_it}E^i\left\{\sum_{j=1}^{p_2}\int_t^{\infty}d\tau\,e^{\beta_i\tau}\|\bar{k}_{ij}(t,\tau)\|^2\cdot\int_t^{\infty}e^{-\beta_is}\|\hat{u}^i(s)\|^2\,ds\right\}\qquad(*)\\
&=\int_0^{\infty}ds\,e^{-\beta_is}E^i\{\|\hat{u}^i(s)\|^2\}\int_0^s e^{-\beta_it}\,dt\int_t^{\infty}e^{\beta_i\tau}\,\mathrm{tr}\,[\bar{K}_i(t,\tau)\bar{K}_i^T(t,\tau)]\,d\tau\\
&=\int_0^{\infty}ds\,e^{-\beta_is}E^i\{\|\hat{u}^i(s)\|^2\}\int_0^s e^{-\beta_it}\bar{a}_i^2(t)\,dt\end{aligned}$$

where we have introduced the term

$$\bar{a}_i^2(t) := \int_t^\infty \operatorname{tr}\,[\bar{K}_i(t,\tau)\bar{K}_i^T(t,\tau)]\, e^{\beta_i \tau}\, d\tau, \qquad t \in [0,\infty). \tag{27}$$

Now, using in (*) the bound on $E^i\{\|\hat{u}^i(s)\|^2\}$ obtained in the proof of Theorem 3.1, we arrive at the result

$$\begin{aligned}\|\bar{\Psi}_i\gamma\|_{\Gamma_i}^2 &\leqq \int_0^\infty ds \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ij}^j \lambda_B^i(s)\, e^{-\beta_i s} E^j\{\|\gamma_j(s,y_j)\|^2\} \int_0^s \bar{a}_i^2(t)\, dt\, e^{-\beta_i t} \\ &= \int_0^\infty ds\, e^{-\beta_i s}\lambda_B^i(s) \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ij}^j\, e^{+\beta_j s}\, e^{-\beta_i s} E^i\{\|\gamma_j(s,y_j)\|^2\} \int_0^s e^{-\beta_i t}\bar{a}_i^2(t)\, dt.\end{aligned}$$

Hence,

$$\begin{aligned}\|\bar{\Psi}\gamma\|_{\Gamma}^2 &= \sum_{i=1}^n \|\bar{\Psi}_i\gamma\|_{\Gamma_i}^2 \\ &\leqq \int_0^\infty \sum_{i\in\mathbb{N}} e^{-\beta_i s} E^i\{\|\gamma_i(s,y_i)\|^2\}\, ds \cdot \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ji}^i \lambda_B^j(s)\, e^{-(\beta_j-\beta_i)s} \int_0^s e^{-\beta_j t}\bar{a}_j^2(t)\, dt \\ &\leqq \bar{\alpha}^2 \|\gamma\|_{\Gamma}\end{aligned}$$

where

$$\bar{\alpha}^2 := \max_{\substack{i,s\\ i\in\mathbb{N}\\ s\in[0,\infty)}} \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ji}^i \lambda_B^j(s)\, e^{-(\beta_j-\beta_i)s} \int_0^s e^{-\beta_j t}\bar{a}_j^2(t)\, dt. \tag{28}$$

Then, the following result readily follows from the classical contraction mapping theorem (cf. [7]).

THEOREM 4.2. *Let $\lambda_B^i(s)$ and $b^i(s)$ be as defined by* (17a)-(17c), *h_{ji}^i be as introduced in Assumption* 2.2, *and $\bar{a}_i^2(\cdot)$ and $\bar{\sigma}^2$ be as defined by* (27)-(28). *Furthermore, let*

$$\bar{\alpha}^2 < 1. \tag{29}$$

Then, for the infinite-horizon stochastic differential game, we have the following:

(i) *There exists a unique Nash equilibrium solution*;

(ii) *The solution is stable under the distributed algorithm* (26).

Remark 4.1. The main condition of the theorem above, i.e., (29), may be overly restrictive, especially if the discount factors in different cost functions are different (i.e., $\beta_i \neq \beta_j$, $i \neq j$). However, if $\beta_i = \beta_j \equiv \beta$, *then* $\bar{\alpha}^2$ can be written as follows:

$$\bar{\alpha}^2 = \max_{\substack{i,s\\ i\in\mathbb{N}\\ s\in[0,\infty)}} \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ji}^i \lambda_B^j(s) \int_0^s e^{-\beta t}\bar{a}_j^2(t)\, dt$$

and if, furthermore,

$$\bar{A}_i^2 := \max_t \bar{a}_i^2(t) < \infty, \qquad i \in \mathbb{N},$$

then condition (29) can be replaced by the courser bound:

$$\frac{1}{\beta}\max_{i,s}\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} h_{ji}^{i}\lambda^{j}{}_{B}(s)\bar{A}_j[1-e^{-\beta s}]<1,$$

which makes the dependence on β more explicit.

To parallel the development in § 3, let us now consider the special class of infinite horizon problems wherein the joint distributions of $x_0, y_1, \cdots, y_n$ are zero-mean Gaussian under all n measures. Let Σ_{0i}, Σ_{ij} be defined as in § 3, and introduce matrix functions $\bar{M}_i$ and $\bar{N}_i$ on $[0,\infty)$, $i\in\mathbb{N}$, satisfying

$$\left(\dot{\bar{M}}_i+\bar{M}_i\left(A-\frac{\beta_i}{2}I\right)+\left(A-\frac{\beta_i}{2}I\right)^{T}\bar{M}_i+Q_i-\bar{M}_iB_iR_{ii}^{-1}B_i^{T}\bar{M}_i\right]\bar{N}_i$$

$$-\sum_{\substack{j\in\mathbb{N}\\ j\neq i}}\bar{M}_iB_jR_{jj}^{-1}B_j^{T}\bar{M}_j\bar{N}_j\Sigma_{ji}=0, \tag{30a}$$

$$\lim_{t_f\to\infty}\bar{M}_i(t_f)=0,$$

$$\dot{\bar{N}}_i=A\bar{N}_i-B_iR_{ii}^{-1}B_i^{T}\bar{M}_i\bar{N}_i-\sum_{\substack{j\in\mathbb{N}\\ j\neq i}}B_jR_{jj}^{-1}B_j^{T}\bar{M}_j\bar{N}_j\Sigma_{ji},\qquad \bar{N}_i(0)=I. \tag{30b}$$

The same reasoning as in the proof of Corollary 3.1, now leads to the following result.

COROLLARY 4.1. *For the infinite horizon differential game problem, let the underlying distributions be zero-mean Gaussian under the n measures. Then, under the conditions of Theorem* 4.1, *we have the following*:

(i) *There exists a unique stable Nash equilibrium solution, under the distributed algorithm* (26), *linear in the measurements*;

(ii) *The unique linear solution is given by*

$$\gamma_i^{N}(t,y_i)=-R_{ii}^{-1}(t)B_i^{T}(t)\bar{M}_i(t)\hat{x}_{it}\quad a.e.\ \mathbf{P}^{i},\quad i\in\mathbb{N},$$

where

$$\hat{x}_{it}=\bar{N}_{it}\Sigma_{0i}y_i,\qquad i\in\mathbb{N},$$

provided that (30) *admits a solution.*

5. Conclusions. In this paper, we have presented results for a class of n-person stochastic linear-quadratic differential games under multiple probabilistic modeling, and with each player acquiring noisy measurement(s) of the initial state. We have seen that if the time interval on which the game is defined is sufficiently short, then the stochastic differential game admits a unique stable Nash equilibrium solution, and this solution can be obtained as the limit of a distributed learning algorithm that involves iteration in the policy space. For the special case when the underlying distributions are Gaussian, this limiting solution is affine in the measurements. If the time horizon is infinite and possibly different discount factors are included in the cost functionals, the same qualitative results hold, this time the conditions of existence and convergence depending also on the magnitude(s) of the discount factor(s).

The distributed algorithm adopted in this paper (see (12) or (26)) allows each player to update on his policy by taking the most recently computed policies of the other players as given. This "parallel scheme" is not the only possible updating scheme if $n\geqq 3$; the players may, for example, update on their policies sequentially either in a fixed order or in an order which changes from stage to stage in a predetermined

manner, or there may be a combination of sequential and parallel schemes, also allowing a delay in the transmission of policy information from one player to others (see [9] and [10] for elucidation of this point in the context of static games). Since the conditions for convergence under these different schemes are, in general, different even in static games [10], [12], it is quite natural to expect them to be different also for the class of stochastic differential games considered in this paper. In principle, the approach developed in this paper could be applied to these other "nonparallel" schemes; however, since the expressions involved are in general quite complicated even for $n=3$, not leading to clean results (see, e.g., [12] for the strictly sequential scheme for the static case and $n=3$), we have not discussed such schemes in this paper.

Another possible major extension would be the study of more general distributed algorithms incorporating memory, such as the relaxation algorithms introduced in [13] for static games. As evidenced from the analysis of [13], incorporation of past values of self-policies in the update mechanism by each player could result in considerable improvements in the conditions of convergence and the speed of convergence towards the unique Nash equilibrium. Developing precise conditions for such relaxation-type algorithms, and obtaining the "best" mix between the other players' past policies and self-policies in the update mechanism would be a promising avenue of research for the future. Yet another challenging extension would be to games with dynamic information patterns, where the players receive on-line (imperfect) information on the current value of the state; but such problems are much more difficult to analyze, even in the case when players have an identical perception of the underlying statistics, partly due to "informational nonuniqueness" (cf. [14]) and partly because a separation of estimation and control is not possible in stochastic games. For the special case of zero-sum games, however, such an extension would be manageable (at least in principle), by taking the results of [15] as the starting point; this is a topic that is currently under investigation.

Appendix.

Proof of Lemma 3.1. First, we note that the zero input solution to (2) is

$$\hat{x}_{it}^{(h)} = X_i(t)\hat{x}_{if} = X_i(t)X_i^{-1}(0)\hat{x}_{i0},$$

$$\lambda_{it}^{(h)} = \Lambda_i(t)\hat{x}_{if} = \Lambda_i(t)X_i^{-1}(0)\hat{x}_{i0}.$$

Now if we can find a particular solution $(\hat{x}_{it}^{(p)}, \lambda_{it}^{(p)})$ to (2), then

$$\hat{x}_{it} = \hat{x}_{it}^{(h)} + \hat{x}_{it}^{(p)}, \qquad \lambda_{it} = \lambda_{it}^{(h)} + \lambda_{it}^{(p)}.$$

To guarantee that $\hat{x}_{it}$ and λ_{it} satisfy the boundary conditions $\hat{x}_{it|_{t=0}} = \hat{x}_{i0}$ and $\lambda_{it|t=t_f} = Q_{if}\hat{x}_{if}$, respectively, let us choose

$$\hat{x}_{it}^{(p)} = \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_0^t F_i(t,\tau)B_j\hat{u}_{j\tau}\,d\tau,$$

$$\lambda_{it}^{(p)} = \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_t^{t_f} G_i(t,\tau)B_j(\tau)\hat{u}_{j\tau}\,d\tau.$$

substituting back into (7), we obtain n sets of equations:

$$
\begin{aligned}
&F_i(t,t)\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} B_j(t)\hat{u}_{jit} + \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_0^t \frac{dF_i(t,\tau)}{dt} B_j(\tau)\hat{u}_{ji\tau}\,d\tau \\
&\quad = A(t)\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_0^t F_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau - B_iR_{ii}^{-1}B_i^T \\
&\qquad \times \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_t^{t_f} G_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau + \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} B_j(t)\hat{u}_{jit}, \\
&\qquad -G_i(t,t)\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} B_j(t)\hat{u}_{jit} + \sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_t^{t_f} \frac{dG_i(t,\tau)}{dt} B_j(\tau)\hat{u}_{ji\tau}\,d\tau \\
&\quad = -Q_i\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_0^t F_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau - A^T(t)\sum_{\substack{j\in\mathbb{N}\\ j\neq i}} \int_t^{t_f} G_i(t,\tau)B_j(\tau)\hat{u}_{ji\tau}\,d\tau.
\end{aligned}
$$

We now observe that the above equations are satisfied if $G_i(t,\tau)$, $F_i(t,\tau)$ satisfy (i) and (ii) of Lemma 3.1, thereby completing the proof. □

REFERENCES

[1] J. C. HARSANYI, *Games with incomplete information played by "Bayesian" players: Part* I. *The basic model*, Management Sci., 14 (1967), pp. 159–182.

[2] ———, *Games with incomplete information played by "Bayesian" players: Part* II. *Bayesian equilibrium points*, Management Sci., 14 (1968), pp. 320–334.

[3] ———, *Games with incomplete information played by "Bayesian" players: Part* III. *The basic probability distribution on the game*, Management Sci., 14 (1968), pp. 486–502.

[4] T. BAŞAR, *An equilibrium theory for multiperson decision making with multiple probabilistic models*, IEEE Trans. Automat. Control, 30 (1985), pp. 118–132.

[5] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

[6] T. BAŞAR AND G. J. OLSDER, *Dynamic Noncooperative Game Theory*, Academic Press, London, New York, 1982.

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

[8] S. LI AND T. BAŞAR, *Distributed learning algorithms for the computation of noncooperative equilibria*, Automatica, 23 (1987), pp. 523–533.

[9] T. BAŞAR, *Stochastic multi-modeling for teams in a game-theoretic framework*, in Optimal Control Theory and Economic Analysis 2, G. Feichtinger, ed., North-Holland, Amsterdam, 1985, pp. 529–548.

[10] T. BAŞAR AND S. LI, *Recursive algorithms for optimal solutions of stochastic teams with decentralized information*, in Proc. 1985 Automatic Control Conference, Boston, MA, June 1985, pp. 231–236.

[11] S. LI, *Recursive schemes for optimal policies in distributed decision making*, M.S. Thesis, University of Illinois, Urbana, IL, 1985.

[12] S. LI AND T. BAŞAR, *Distributed learning schemes for Nash games*, in Proc. IFAC Workshop on Modelling, Decision and Games with Applications to Social Phenomena, Beijing, August 1986, pp. 307–316.

[13] T. BAŞAR, *Relaxation techniques and asynchronous algorithms for on-line computation of noncooperative equilibria*, J. Economic Dynamics Control, 11 (1987), pp. 531–549.

[14] ———, *Informationally nonunique equilibrium solutions in differential games*, SIAM J. Control Optim., 15 (1977), 636–660.

[15] ———, *On the saddle-point solutions of a class of stochastic differential games*, J. Optim. Theory Appl., 33 (1981), pp. 539–556.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 579-607, May 1989

008

SOME ESTIMATES FOR FINITE DIFFERENCE APPROXIMATIONS*

JOSE-LUIS MENALDI†

Abstract. Some estimates for the approximation of optimal stochastic control problems by discrete time problems are obtained. In particular an estimate for the solutions of the continuous time versus the discrete time Hamilton-Jacobi-Bellman equations is given. The technique used is more analytic than probabilistic.

Key words. diffusion process, Markov chain, dynamic programming, finite difference, Hamilton-Jacobi-Bellman equations

AMS(MOS) subject classifications. 65K10, 65G99, 49D25, 93E25, 93E20

Introduction. We are interested in the approximation of optimal control problems for diffusion processes by means of finite difference methods. It is well known (e.g., Kushner [16], [17]) that a basic probabilistic counterpart is the approximation of a diffusion process by a Markov chain. A typical problem in stochastic control theory is the following.

In a complete filtered probability space $(\Omega, P, \mathcal{F}, \mathcal{F}(t), t \geqq 0)$ suppose we have two progressively measurable processes $(y(t), \lambda(t), t \geqq 0)$ satisfying the following stochastic differential equation in the Itô sense:

$$\begin{aligned} &dy(t) = g(y(t), \lambda(t))\, dt + \sigma(y(t), \lambda(t))\, dw(t), \qquad t \geqq 0, \\ &y(0) = x, \end{aligned} \tag{0.1}$$

for given x, g, σ, and some n-dimensional Wiener process $(w(t), t \geqq 0)$. The processes $(y(t), t \geqq 0)$ and $(\lambda(t), t \geqq 0)$ represent the state in $\mathcal{R}^d$ and the control in Λ (a compact metric space) of the dynamic system, respectively.

The cost functional is given by

$$J(x, \lambda) = E\left\{\int_0^{\tau} f(y(t), \lambda(t))\, e^{-\alpha t}\, dt\right\}, \tag{0.2}$$

where f is a given function, $\alpha > 0$, and τ is the first exit time of a domain D in $\mathcal{R}^d$ for the process $(y(t), t \geqq 0)$.

The associated Hamilton-Jacobi-Bellman (HJB) equation (e.g., Bensoussan and Lions [2], Fleming and Rishel [9], Krylov [14]) to be satisfied by the optimal cost

$$u(x) = \inf\{J(x, \lambda): \text{any control } \lambda(\cdot)\} \tag{0.3}$$

is indeed

$$\begin{aligned} &\alpha u = \inf\{L(\lambda)u + f(\cdot, \lambda): \lambda \in \Lambda\} \quad \text{in } D, \\ &u = 0 \quad \text{on } \in \partial D, \end{aligned} \tag{0.4}$$

with the differential operator

$$L(\lambda) = \frac{1}{2} \sum_{i,j=1}^{d} \left(\sum_{k=1}^{n} \sigma_{ik}(\cdot, \lambda)\sigma_{jk}(\cdot, \lambda) \right) \partial_{ij} + \sum_{i=1}^{d} g_i(\cdot, \lambda)\, \partial_i, \tag{0.5}$$

where $\partial_{ij}, \partial_i$ denote the partial derivatives and $g = (g_i, i = 1, \cdots, d)$, $\sigma = (\sigma_{ik}, i = 1, \cdots, d, k = 1, \cdots, n)$.

* Received by the editors September 8, 1986; accepted for publication (in revised form) July 12, 1988. This research was partly supported by National Science Foundation grant DMS-8702236.

† Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

Let $\mathcal{R}_h^d$ denote a h-finite difference grid in $\mathcal{R}^d$. Consider the finite difference operator

$$L_h(\lambda)\varphi(x) = h^{-1} \sum_{k=1}^{n} \{\beta_k^+(x, \lambda, h)[\varphi(x + \gamma_k^+(x, \lambda, h)) - \varphi(x)] + \beta_k^-(x, \lambda, h)[\varphi(x + \gamma_k^-(x, \lambda, h)) - \varphi(x)]\}, \tag{0.6}$$

where the coefficients satisfy

$$\beta_k^{\pm}(x, \lambda, h) \geqq 0 \quad \forall x, \lambda, h,$$
$$x + \gamma_k^{\pm}(x, \lambda, h) \in \mathcal{R}_h^d \quad \forall x \in \mathcal{R}_h^d, \quad \lambda \in \Lambda. \tag{0.7}$$

The finite difference approximation of the HJB equation (0.4) using the operator (0.6) is

$$\alpha u_h = \inf\{L_h(\lambda)u_h + f(\cdot, \lambda) : \lambda \in \Lambda\} \quad \text{in } D_h,$$
$$u_h = 0 \quad \text{in } \mathcal{R}_h^d \setminus D_h, \tag{0.8}$$

where D_h is the set of points in $\mathcal{R}_h^d$ belonging to D.

Our purpose is to estimate the difference

$$\sup\{|u(x) - u_h(x)| : x \in D_h\} \tag{0.9}$$

in terms of the parameter h. We expect to dominate (0.9) by

$$\sup\{\inf\{|l(x, \lambda) - l(x', \lambda)| : x' \in \mathcal{R}_h^d\} : x \in \mathcal{R}^d, \lambda \in \Lambda\}, \quad \text{for } l = f, g_i, \sigma_{ik},\ i = 1, \cdots, d,\ k = 1, \cdots, n. \tag{0.10}$$

For instance, if f, g, σ are Lipschitz-continuous in x, then

$$|u(x) - u_h(x)| \leqq Ch^{1/2} \quad \forall x \in D_h, \quad h \in (0, 1], \tag{0.11}$$

for some constant C independent of x and h.

Let us mention that finite difference operators of the form (0.6) satisfy automatically the so-called discrete maximum principle. Problem (0.8) is indeed the discrete HJB equation associated with some suitable optimal control problem of a Markov chain. We remark that several computational methods are available for the discrete HJB equation (0.8) (e.g., Kushner and Kleinman [18], Puterman [29], Puterman and Brumelle [30], Quadrat [31], and Theosys [33]).

Actually, the objective of the paper is to show how the underlying technique can be used with a typical problem (0.1)–(0.5). The probabilistic interpretation of the finite difference operator (0.6) is part of the key idea. From a purely stochastic control viewpoint, an estimate on an approximation to the optimal cost is certainly of great value. However, we may question how optimal the discrete optimal feedback is when it is applied to the actual continuous time problem. Toward an answer to the preceding questions, we can argue in the following way. First of all, what really matters for the optimizers is to know how far they are from the minimum cost in the real model. The stochastic equation (0.1) is only an approximation of the real evolution, as well as being the Markov chain associated with the operator (0.6). Our claim is that by preserving the structure of the problem, i.e., to have a probabilistic interpretation of the approximating HJB equation (0.8), and by getting some estimates of the convergence of the corresponding optimal costs, we cannot be far away from the real model.

Even if the Markov chain associated with the operator (0.6) always has finite state, we may want to discretize the set Λ, just to improve the implementation of the

infimum in equation (0.8). In this case, we can replace Λ in (0.7) and (0.8) by a discretization $\Lambda(h)$, and similar results hold true (cf. [24]).

Deterministic versions along with the same kind of ideas can be found in Capuzzo-Dolcetta [4], Capuzzo-Dolcetta and Ishii [5], Crandall and Lions [6], Falcone [8], Gonzalez and Rofman [12], Menaldi and Rofman [27], and Souganidis [36].

The cases where the discount factor α is actually a function, the coefficients g, σ are time-dependent, the horizon is finite, the HJB equation is indeed a set of inequalities, and the domain D is unbounded can also be studied.

In § 1 we consider the one-dimensional case. Even if this case is very restrictive, we obtain enough information from it to deal with the multidimensional case. Moreover, this section can stand by itself, but we believe it is a natural step in the technique to be developed. General problems are treated in § 2.

1. One-dimensional case. It is clear that for one-dimensional problems we dispose of many classic tricks, probably more efficient in practice than the one we will describe. However, we claim that by studying this simple case we may obtain some nonstandard ways of looking at a multidimensional finite difference scheme.

Let g, σ be real continuous functions on $\mathcal{R} \times \Lambda$ such that

$$\begin{aligned} &|g(x,\lambda)|+|\sigma(x,\lambda)| \leqq C \quad \forall x \in \mathcal{R}, \quad \lambda \in \Lambda, \\ &|g(x,\lambda)-g(x',\lambda)|+|\sigma(x,\lambda)-\sigma(x',\lambda)| \leqq K|x-x'| \quad \forall x, x' \in \mathcal{R}, \quad \lambda \in \Lambda, \end{aligned} \tag{1.1}$$

for some constants $C = C(g, \sigma)$ and $K = K(g, \sigma)$. The set Λ is a compact metric space, generally a compact subset of $\mathcal{R}^m$.

On a complete Wiener space $(\Omega, P, \mathcal{F}, \mathcal{F}(t), w(t), t \geqq 0)$, i.e., $(\Omega, P, \mathcal{F})$ is a complete probability space, $(\mathcal{F}(t), t \geqq 0)$ is a right-continuous family of complete sub-σ-algebras of $\mathcal{F}$, $(w(t), t \geqq 0)$ is a one-dimensional standard Wiener process adapted to $(\mathcal{F}(t), t \geqq 0)$; we consider the controlled diffusion process

$$dy(t) = g(y(t), \lambda(t))\, dt + \sigma(y(t), \lambda(t))\, dw(t), \qquad t > 0, \tag{1.2}$$

where the control $(\lambda(t), t \geqq 0)$ is a progressively measurable process taking values in Λ. Its associated infinitesimal generator $L(\lambda)$ is the second-order differential operator

$$L(\lambda)\varphi = \tfrac{1}{2}\sigma^2(\cdot, \lambda)\varphi'' + g(\cdot, \lambda)\varphi', \qquad \sigma(\cdot, \cdot) \geqq 0, \tag{1.3}$$

where φ' and φ'' are the first and second derivatives of φ.

For the moment, let us forget about the h-finite grid $\mathcal{R}_h$, i.e., the last condition of (0.7) is disregarded. Consider the finite difference operator

$$\begin{aligned} &L_h(\lambda)\varphi = \frac{1}{h}\left[\frac{1}{2}\varphi(\cdot + gh + \sigma\gamma\sqrt{h}) + \frac{1}{2}\varphi(\cdot + gh - \sigma\gamma\sqrt{h}) - \varphi\right], \\ &g = g(x,\lambda), \quad \sigma = \sigma(x,\lambda), \quad \gamma = \gamma(x,\lambda,h); \end{aligned} \tag{1.4}$$

the function $\gamma \geqq 0$ is to be chosen later (cf. (1.8)).

In § 1.1 we will construct a controlled Markov chain associated with the finite difference operator, from which a piecewise constant (on stochastic time intervals) process $(y_h(t); t \geqq 0)$ is defined in such a way that

$$E \sup\{|y(t) - y_h(t)|^p e^{-\alpha t} : t \geqq 0\} \leqq Ch^{p/2} \quad \forall h \in (0, 1], \tag{1.5}$$

for some constants C, $\alpha > 0$ depending only on g, σ, and $p > 0$ uniformly with respect to a class of controls to be specified.

Next, we use this estimate to obtain (0.11) for a linear equation, i.e., without control λ.

In § 1.3 we realize the above technique gives only a one-sided estimate of the type (0.11) for nonlinear problems. The difficulty is the lack of information on the optimal control $\lambda(\cdot)$. At this point, we need to use analytic techniques to obtain (0.11).

1.1. A Markov chain. Define

$$\begin{aligned}
&\tau(x,\lambda,h,w)=\inf\{t\geqq 0:\ g(x,\lambda)(t-h)+\sigma(x,\lambda)w(t)\\
&\qquad\qquad\qquad\qquad\text{equals either } \delta(x,\lambda,h) \text{ or } -\delta(x,\lambda,h)\},\\
&\delta(x,\lambda,h)=\sigma(x,\lambda)\gamma(x,\lambda,h)\sqrt{h}, \qquad w(0)=0,\\
&\xi(x,\lambda,h,w)=g(x,\lambda)\tau(x,\lambda,h,w)+\sigma(x,\lambda)w(\tau(x,\lambda,h,w)).
\end{aligned} \tag{1.6}$$

Note that $w(\cdot)$ is a standard Wiener process and $\tau=h$ and $\xi=gh$ whenever σ vanishes.

Let $\lambda(\cdot)$ be a feedback control, i.e., a Borel-measurable function from $\mathcal{R}$ into Λ. We construct by induction the sequences of random variables $(X_n, \theta_n, n=0,1,\cdots)$ as follows. For a given initial data x,

$$\begin{aligned}
&X_0=x, \quad \theta_0=0, \quad w_0(t)=w(t),\\
&X_{n+1}=X_n+\xi(X_n,\lambda(X_n),h,w_n),\\
&\theta_{n+1}=\theta_n+\tau(X_n,\lambda(X_n),h,w_n),\\
&w_{n+1}(t)=w(t+\theta_n)-w(\theta_n), \qquad n=0,1,\cdots.
\end{aligned} \tag{1.7}$$

If instead of a feedback control $\lambda(\cdot)$ we have a nonanticipating control $(\lambda_n, n=0,1,\cdots)$, where λ_n is a random variable valued in Λ and adapted to $(X_0,\cdots,X_{n-1})$, then the procedure (1.7) still works.

Let us define the function $\gamma(x,\lambda,h)$ by

$$\begin{aligned}
&\gamma=0 \quad \text{if } 0\leqq\sigma<|g|\sqrt{h},\\
&\gamma=1 \quad \text{if } g=0,\\
&\gamma=\gamma_0(g\sigma^{-1}\sqrt{h}) \quad \text{if } 0<|g|\sqrt{h}\leqq\sigma,
\end{aligned} \tag{1.8}$$

where

$$\gamma_0(r)=(2r)^{-1}\ln[e^{2r^2}+\operatorname{sign}(r)(e^{4r^2}-1)^{1/2}], \tag{1.9}$$

for $r\neq 0$, $-1\leqq r\leqq 1$. Note that $\gamma_0(r)>0$; moreover,

$$0<\gamma_0(r)-1\leqq|r| \quad \forall r\in[-1,0)\cup(0,1]. \tag{1.10}$$

This implies the inequality

$$|\sigma(x,\lambda)\gamma(x,\lambda,h)-\sigma(x,\lambda)|\leqq 2|g(x,\lambda)|\sqrt{h}, \tag{1.11}$$

for every x, λ, h.

THEOREM 1.1. *If we choose $\gamma(x,\lambda,h)$ by* (1.8), *then for any feedback $\lambda(\cdot)$ the procedure* (1.7) *defines a Markov chain $(X_n, n=0,1,\cdots)$ with transition probability determined by*

$$\begin{aligned}
&E(\varphi(X_{n+1})\,|\,X_n=x)=\Pi_h(\lambda(x))\varphi(x),\\
&\Pi(\lambda)\varphi(x)=\tfrac{1}{2}\varphi(x+g(x,\lambda)h+\sigma(x,\lambda)\gamma(x,\lambda,h)\sqrt{h})\\
&\qquad\qquad+\tfrac{1}{2}\varphi(x+g(x,\lambda)h-\sigma(x,\lambda)\gamma(x,\lambda,h)\sqrt{h}),
\end{aligned} \tag{1.12}$$

and a sequence $(\theta_n, n=0,1,\cdots)$ of stopping times relative to $(\mathcal{F}(t), t\geqq 0)$, with independent increments $\tau_n=\theta_n-\theta_{n-1}$,

$$E\tau_n=h \quad \forall n=1,2,\cdots. \tag{1.13}$$

Proof. Without loss of generality, we may assume g and σ constants. Consider the two functions $u(x)$ and $v(x)$ defined by the equations

$$\tfrac{1}{2}\sigma^2 u'' + gu' = -1 \quad \text{in } (\delta, \delta), \quad u(-\delta) = u(\delta) = 0,$$

$$\tfrac{1}{2}\sigma^2 v'' + gv' = 0 \quad \text{in } (-\delta, \delta), \quad v(-\delta) = 0, \quad v(\delta) = 1,$$

where $\delta = \sigma\gamma\sqrt{h}$. If $-\delta \leqq -gh \leqq \delta$, then

$$E\tau(h, \cdot) = u(-gh), \qquad P(\xi(h, \cdot) = gh + \delta) = v(-gh), \tag{1.14}$$

with the notation (1.6). Since we can compute explicitly,

$$u(x) = -\frac{x}{g} + \frac{\delta}{g}\left(\frac{e^{\alpha\delta} + e^{-\alpha\delta} - 2e^{\alpha x}}{e^{\alpha\delta} - e^{-\alpha\delta}}\right), \quad \alpha = 2g\sigma^{-2}, \quad v(x) = \frac{e^{\alpha\delta} - e^{-\alpha x}}{e^{\alpha\delta} - e^{-\alpha\delta}},$$

yielding

$$u(-gh) = h + \frac{\gamma h}{r}\left(\frac{e^{2r\gamma} + e^{-2r\gamma} - 2e^{2r^2}}{e^{2r\gamma} - e^{-2r\gamma}}\right), \quad v(-gh) = \frac{e^{2r\gamma} - e^{2r^2}}{e^{2r\gamma} - e^{-2r\gamma}}, \quad r = g\sigma^{-1}\sqrt{h}.$$

Suppose we have chosen γ such that $u(-gh) = h$, i.e.,

$$e^{2r\gamma} + e^{-2r\gamma} - 2e^{2r^2} = 0, \qquad \gamma > 1. \tag{1.15}$$

Since

$$1 - v(-gh) = \frac{e^{2r^2} - e^{2r\gamma}}{e^{2r\gamma} - e^{-2r\gamma}},$$

the relation (1.15) implies $v(-gh) = \frac{1}{2}$, i.e.,

$$E\tau(h, \cdot) = h, \qquad P(\xi(h, \cdot) = gh \pm \sigma\gamma\sqrt{h}) = \tfrac{1}{2}, \tag{1.16}$$

whenever $0 < |g|\sqrt{h} \leqq \sigma\gamma$. Note that (1.16) still holds if we take $\gamma = 1$ for $g = 0$ and that (1.15) gives $\gamma = \gamma_0(r)$ as in (1.9), for $0 < |g|\sqrt{h} \leqq \sigma$, because $\gamma > 1$.

If $0 \leqq \sigma < |g|\sqrt{h}$, then the equalities (1.14) hold true for functions u and v satisfying

$$\tfrac{1}{2}\sigma^2 u'' + gu' = -1 \quad \text{in } (-\infty, -\delta), \quad u(-\delta) = 0, \quad u \text{ with polynomial growth},$$

$$\tfrac{1}{2}\sigma^2 v'' + gv' = 0 \quad \text{in } (-\infty, -\delta), \quad v(-\delta) = 1, \quad v \text{ with polynomial growth},$$

for $g > 0$ and replacing the interval $(-\infty, -\delta)$ by $(\delta, +\infty)$ if $g < 0$. It is clear that $v = 1$ and

$$u(x) = -\frac{x}{g} + \frac{\delta}{|g|}.$$

So

$$u(-gh) = h + |g|^{-1}\sigma\gamma\sqrt{h},$$

and $\gamma = 0$ is the right choice.

All of the above proves (1.12) and (1.13) after using standard facts on Brownian motions. □

Remark 1.1. If $g = 0$ and $\sigma = 1$ then the construction (1.7) coincides with the classic Skorokhod's representation (cf. [35]).

For a given feedback $\lambda(\cdot)$ let us denote by $(y_x^h(t, \lambda(\cdot)), t \geqq 0)$ and $(\lambda^h(t), t \geqq 0)$ the processes

$$
\begin{aligned}
&y_x^h(t, \lambda(\cdot)) = X_n \quad \text{if } \theta_n \leqq t < \theta_{n+1}, \quad n = 0, 1, \cdots, \\
&\lambda^h(t) = \lambda(X_n) \quad \text{if } \theta_n \leqq t < \theta_{n+1}, \quad n = 0, 1, \cdots,
\end{aligned} \tag{1.17}
$$

where $(X_n, \theta_n, n = 0, 1, \cdots)$ are defined by (1.7) with the choice (1.8) of function γ.

Note that these processes are adapted to $(\mathcal{F}(t), t \geqq 0)$ and piecewise constants on stochastic intervals. This approach is different from the one used by Pardoux and Talay [28]. Our partition is on the range, i.e., X_n takes values in a variable grid of $\mathcal{R}$, and the time partition is chosen accordingly; our time intervals are random.

Now consider the controlled diffusion process $(y(t) = y_x(t, \lambda^h), t \geqq 0)$ given by the stochastic equation (1.2) with initial data $y(0) = x$ and control process $\lambda(t) = \lambda^h(t)$.

THEOREM 1.2. *Let the assumption* (1.1) *and the choice* (1.8) *hold. Then for any positive number p there exist two positive constants C, α depending only on p and the constants* $C(g, \sigma)$, $K(g, \sigma)$ *of* (1.1) *such that*

$$
E \sup \{|y_x(t, \lambda^h) - y_x^h(t, \lambda(\cdot))|^p e^{-\alpha t} : t \geqq 0\} \leqq C h^{p/2}, \tag{1.18}
$$

uniformly for any feedback $\lambda(\cdot)$ *and x in* $\mathcal{R}$.

Proof. Based on the procedure (1.6), (1.7) we have

$$
X_{n+1} = X_n + g(X_n, \lambda(X_n))(\theta_{n+1} - \theta_n) + \sigma(X_n, \lambda)(w(\theta_{n+1}) - w(\theta_n)), \qquad n = 0, 1, \cdots,
$$

which gives

$$
X_n = x + \int_0^{\theta_n} g(y^h(t), \lambda^h(t))\, dt + \int_0^{\theta_n} \sigma(y^h(t), \lambda^h(t))\, dw(t),
$$

where $y^h(t) = y_x^h(t, \lambda(\cdot))$ and $\lambda^h(t)$ are the processes defined by (1.17). If we set

$$
q^h(t) = x + \int_0^t g(y^h(s), \lambda^h(s))\, ds + \int_0^t \sigma(y^h(s))\, dw(s), \qquad t \geqq 0,
$$

then

$$
q^h(t) - y^h(t) = g(X_n, \lambda(X_n))(t - \theta_n) + \sigma(X_n, \lambda(X_n))(w(t) - w(\theta_n)) \quad \text{if } \theta_n \leqq t < \theta_{n+1}.
$$

Again, in view of the definition (1.6) we deduce

$$
|q^h(t) - y^h(t)| \leqq C(g, \sigma)\sqrt{h} \quad \forall t \geqq 0, \quad 0 < h \leqq 1, \tag{1.19}
$$

where $C(g, \sigma)$ is the constant of the assumption (1.1).

Now, consider the process $z(t) = y(t) - q^h(t)$, with $y(t) = y_x(t, \lambda^h)$,

$$
\begin{aligned}
dz(t) = {} & [g(y(t), \lambda^h(t)) - g(y^h(t), \lambda^x(t))]\, dt \\
& + [\sigma(y(t), \lambda^h(t)) - \sigma(y^h(t), \lambda^h(t))]\, dw(t), \qquad t \geqq 0, \quad z(0) = 0,
\end{aligned}
$$

and apply Itô's formula to the function

$$
\varphi(z, t) = (\beta^2 + z^2)^{p/2} e^{-\alpha t}, \qquad \alpha, \beta, p > 0,
$$

to get

$$
\begin{aligned}
d\varphi(z(t), t) = {} & \{pz(t)[g(y(t), \lambda^h(t)) - g(y^h(t), \lambda^h(t))](\beta^2 + z^2(t)) \\
& + \tfrac{1}{2}(p\beta^2 + p(p-1)z^2(t))[\sigma(y(t), \lambda^h(t)) - \sigma(y^h(t), \lambda^h(t))]^2 \\
& - \alpha(\beta^2 + z^2(t))^2\}(\beta^2 + z^2(t))^{p/2-2} e^{-\alpha t}\, dt \\
& + pz(t)[\sigma(y(t), \lambda^h(t)) - \sigma(y^h(t), \lambda^h(t))](\beta^2 + z^2(t))^{p/2-1} e^{-\alpha t}\, dw(t).
\end{aligned}
$$

If we take $\alpha > \alpha_p$,

$$\alpha_p = p \sup \{[g(x,\lambda)-g(x',\lambda)](x-x')^{-1} + (p \vee 1)[\sigma(x,\lambda)-\sigma(x',\lambda)]^2(x-x')^{-2}: x \neq x' \text{ in } \mathcal{R},\ \lambda \text{ in } \Lambda\}, \tag{1.20}$$

where $\vee$ denotes the maximum, and

$$\beta = \sup \{|q^h(t)-y^h(t)|: t \geqq 0,\ \omega \text{ in } \Omega\};$$

then we obtain

$$d\varphi(z(t),t) \leqq (\alpha_p - \alpha)\varphi(z(t),t)\,dt + dM_t, \qquad t \geqq 0,$$
$$|y(t)-y^h(t)|^p e^{-\alpha t} \leqq \varphi(z(t),t), \qquad t \geqq 0,$$

with M_t being the martingale term. In virtue of (1.19) we deduce

$$E\left\{|y(t)-y^h(t)|^p e^{-\alpha t} + (\alpha-\alpha_p)\int_0^t |y(s)-y^h(s)|^p e^{-\alpha s}\,ds\right\} \leqq [C(g,\sigma)\sqrt{h}]^p \quad \forall t \geqq 0, \quad 0 < h \leqq 1. \tag{1.21}$$

Next, by means of the stochastic inequality

$$E \sup\left\{\left|\int_0^t \varphi(s)\,dw(s)\right|: t \geqq 0\right\} \leqq 3E\left\{\left(\int_0^\infty \varphi^2(t)\,dt\right)^{1/2}\right\},$$

we bound the martingale term

$$E \sup\left\{\left|\int_0^t dM_s\right|: t \geqq 0\right\} \leqq 3pK(g,\sigma)\left(E\left\{\int_0^\infty \varphi^2(z(t),t)\,dt\right\}\right)^{1/2},$$
$$\{\varphi^2(z(t),t)\} \leqq [C(g,\sigma)\sqrt{h}]^{2p} e^{-(2\alpha-\alpha_{2p})t}, \quad t \geqq 0.$$

Hence we obtain (1.18) for $2\alpha > \alpha_{2p}$ as in (1.20) and

$$C = C(g,\sigma)[1+3pK(g,\sigma)(2\alpha-\alpha_{2p})^{-1}],$$

where $C(g,\sigma)$ and $K(g,\sigma)$ are the constants of (1.1). □

Remark 1.2. Notice that the constant α_p defined by (1.20) is bounded by $p(p \vee 1)K(g,\sigma)$, the constant of (1.1). It is clear then that α_p vanishes as p goes to zero.

Remark 1.3. Similarly, we can show for any $t \geqq 0$ the estimate

$$E\{|y_x^h(t,\lambda(\cdot)) - y_{x'}^h(t,\lambda(\cdot))|^p e^{-\alpha_p t}\} \leqq (C^2(g,\sigma)h + |x-x'|^2)^{p/2}, \tag{1.22}$$

where $C(g,\sigma)$, α_p are the constants of (1.1), (1.20) and x, x' belong to $\mathcal{R}$, h in $(0,1]$, and the feedback $\lambda(\cdot)$ is arbitrary.

1.2. The linear equation. In this section we consider the case without controls, i.e., the set Λ reduces to one element, and we drop it.

Recall the stochastic differential equation

$$dy(t) = g(y(t))+\sigma(y(t))\,dw(t), \qquad t \geqq 0, \quad y(0) = x, \tag{1.23}$$

where g, σ are bounded and Lipschitz-continuous. For a bounded and uniformly continuous function f we set

$$u(x,t) = E\left\{\int_0^t f(y_x(s))\,ds\right\}, \quad \forall x \in \mathcal{R}, \quad t \geqq 0. \tag{1.24}$$

This function is the unique solution of the following partial differential equation that is bounded and uniformly continuous:

$$\partial_t u(x,t) = Lu(x,t) + f(x) \quad \forall x \in \mathcal{R}, \quad t > 0, \quad u(x,0) = 0 \quad \forall x \in \mathcal{R} \tag{1.25}$$

where ∂_t denotes the partial derivative in t and L is the differential operator

$$L\varphi(x,t) = \tfrac{1}{2}\sigma^2(x)\partial_x^2\varphi(x,t) + g(x)\partial_x\varphi(x,t). \tag{1.26}$$

The partial differential equation (PDE) (1.25) is understood in the Schwartz' distributions sense. On the other hand, we set

$$u_h(x, nh) = h \sum_{i=0}^{n} E\{f(X_i)\}, \qquad n = 0, 1, \cdots,$$

where $(X_n, \theta_n, n = 0, 1, \cdots)$ is the sequence of Theorem 1.1. It is clear that

$$u_h(x, nh) = E\left\{\int_0^{\theta_n} f(y_x^h(t))\, dt\right\}, \qquad n = 0, 1, \cdots, \tag{1.27}$$

with the notation of Theorem 1.2. For convenience we set

$$u_h(x,t) = u_h(x, nh) \quad \text{if } nh \leqq t < (n+1)h.$$

By means of Markov's property, we can deduce

$$\begin{aligned} &\nabla_h u_h(x,t) = L_h u_h(x,t) + f(x) \quad \forall x \in \mathcal{R}, \quad t > 0, \\ &u_h(x,0) = 0 \quad \forall x \in \mathcal{R}, \end{aligned} \tag{1.28}$$

where L_h is now the finite difference operator

$$\begin{aligned} L_h\varphi(x) = \frac{1}{2h}[&\varphi(x + g(x)h + \sigma(x)\gamma(x,h)\sqrt{h}) \\ &+ \varphi(x + g(x)h - \sigma(x)\gamma(x,h)\sqrt{h}) - 2\varphi(x)] \quad \forall h \text{ in } (0,1], \end{aligned} \tag{1.29}$$

and

$$\nabla_h\varphi(t) = \frac{1}{h}[\varphi(t+h) - \varphi(t)] \quad \forall h \text{ in } (0,1]. \tag{1.30}$$

Note that x belongs to $\mathcal{R}$, so our Markov chain has states in $\mathcal{R}$. Actually, we discretize first the time variable and then the state variable.

Denote by $\rho(r)$ the modulus of continuity of f, i.e.,

$$\rho(r) = \sup\{|f(x) - f(x')| : x, x' \in \mathcal{R}, |x - x'| \leqq r\}. \tag{1.31}$$

THEOREM 1.3. *Under the assumptions of Theorem* 1.2, *for any* p, $T > 0$ *there exists a constant* C *depending only on* p, T, *the Lipschitz constants of* g, σ, *and the bound of* f, *such that*

$$|u(x,t) - u_h(x,t)| \leqq C[\sqrt{h} + \rho(r) + (r^{-1}\sqrt{h})^p] \quad \forall r > 0, \tag{1.32}$$

valid for any x *in* $\mathcal{R}$, t *in* $[0, T]$, h *in* $(0, 1]$.

Proof. In view of the representations (1.24) and (1.27) we have

$$|u(x,t) - u_h(x,t)| \leqq E\left\{\int_0^t |f(y(s)) - f(y^h(s))|\, ds\right\} + C(f)E\{|\theta_n - t|\} = I + II,$$

where

$$|f(x)-f(x')|\leqq C(f) \quad \forall x, x'\in \mathcal{R}.$$

Thus,

$$I\leqq \alpha^{-1}(e^{\alpha t}-1)C(f)r^{-p}E\left\{\int_0^t |y(s)-y^h(s)|^p\, e^{-\alpha s}\, ds\right\}$$
$$+ t\rho(r)\leqq \alpha^{-1}(e^{\alpha T}-1)C(f)C_1(r^{-1}\sqrt{h})^p + T\rho(r),$$

with α, $C=C_1$ being the constants of (1.18) in Theorem 1.2. Also,

$$E\{|\theta_n - t|\}\leqq h + E\{|\theta_n - nh|\}\leqq h + (E\{|(\tau_1-h)+\cdots+(\tau_n-h)|^2\})^{1/2}$$
$$= h + (nE\{(\tau_1-h)^2\})^{1/2}\leqq h + C_2\sqrt{th},$$

where $\tau_i=\theta_i-\theta_{i-1}$, and C_2 is a constant such that

$$E\{(\tau_1-h)^2\}\leqq (C_2h)^2 \quad \forall h\in(0,1]. \tag{1.33}$$

It is clear that the above proves (1.32) provided we have established (1.33).

To show (1.33), we see that if $-\delta\leqq -gh\leqq\delta$, $\delta=\sigma\gamma\sqrt{h}$ then the characteristic function of τ_1,

$$u(x,s)=E\{e^{s\tau_1}\}, \qquad s>0 \text{ fixed},$$

is the solution of the differential equation

$$\tfrac{1}{2}\sigma^2u''+gu'-su=0 \quad \text{in } (-\delta,\delta), \quad u(-\delta,s)=u(-\delta,s)=1,$$

and

$$E\{(\tau_1)^2\}=\frac{\partial^2 u}{\partial s}(-gh,0).$$

Hence, after some calculations we obtain (1.33). □

Remark 1.4. Analogously to the above theorem, and by means of Remark 1.3, we can prove that

$$|u_h(x,t)-u_h(x',t)|\leqq C\{\sqrt{h}+\rho(r)+r^{-p}[h+|x-x'|^2]^{p/2}\} \quad \forall r>0 \tag{1.34}$$

for any h in $(0,1]$, x in $\mathcal{R}$, t in $[0,T]$ and some constant C depending only on the bound of f, the Lipschitz constants of g, σ and the constants T, $p>0$. Actually, we can do better, i.e., in the estimates (1.22) and (1.34) we may have the right-hand side with $h=0$, but this requires the use of another explicit Markov chain, the one used in § 1.3.

1.3. Fully nonlinear equation. Let us return to the control problem (0.1)-(0.5) for one dimension, i.e., D is the whole real line $\mathcal{R}$, Λ is some compact subset of $\mathcal{R}$, $n=d=1$ in (0.5). Recall that for any adapted control process $(\lambda(t), t\geqq 0)$ we obtain the state process $(y(t)=y_x(t,\lambda), t\geqq 0)$ as the solution of the stochastic differential equation (1.2) with initial condition $y(0)=x$. Next, the cost functional is defined by

$$J(x,\lambda)=E\left\{\int_0^\infty f(y(t),\lambda(t))\, e^{-\alpha t}\, dt\right\}, \tag{1.35}$$

and the optimal cost is

$$u(x)=\inf\{J(x,\lambda): \lambda \text{ any control process}\}. \tag{1.36}$$

The associated HJB equation is

$$\alpha u = \inf\{L(\lambda)u + f(\cdot, \lambda): \lambda \in \Lambda\} \quad \text{in } \mathcal{R}, \tag{1.37}$$

which is indeed an ordinary differential equation in the real line, since $L(\lambda)$ is given by (1.3). If the data are smooth and the operator is uniformly elliptic, then the HJB equation (1.37) has one and only one solution with Lipschitz second derivative (cf. Krylov [13]). In general we use either the viscosity solution (cf. Lions [20]) or the maximum subsolution in the Schwartz' distribution sense (cf. Lions and Menaldi [21]).

The approximate control is then

$$\begin{aligned} &J_h(x, \lambda(\cdot)) = E\left\{\int_0^\infty f(y^h(t), \lambda(y^h(t)))\chi_h^\alpha(t)\, dt\right\}, \\ &\chi_h^\alpha(t) = (1+\alpha h)^{-n} \quad \text{if } \theta_n \leqq t < \theta_{n+1}, \quad n = 0, 1, \cdots, \end{aligned} \tag{1.38}$$

where $(y^h(t) = y_x^h(t, \lambda(\cdot)), t \geqq 0)$ and $(\theta_n, n = 0, 1, \cdots)$ are defined by (1.17) and (1.7). Note that

$$J_h(x, \lambda(\cdot)) = E\left\{h \sum_{n=0}^{\infty} f(X_n, \lambda(X_n))(1+\alpha+\alpha h)^{-n}\right\}. \tag{1.39}$$

The optimal cost is

$$u_h(x) = \inf\{J_h(x, \lambda(\cdot)): \lambda(\cdot) \text{ feedback control}\}, \tag{1.40}$$

$$\alpha u_h = \inf\{L_h(\lambda)u_h + f(\cdot, \lambda): \lambda \in \Lambda\} \quad \text{in } \mathcal{R}. \tag{1.41}$$

It is clear that an estimate of the type (1.18) will provide only a one-side bound for the rate of convergence of u_h toward u. Then we will modify the continuous time control problem as follows.

To simplify the exposition we assume g, σ Lipschitz-continuous in the control variable, i.e.,

$$|g(x, \lambda) - g(x, \lambda')| + |\sigma(x, \lambda) - \sigma(x, \lambda')| \leqq K|\lambda - \lambda'| \quad \forall x \in \mathcal{R}, \lambda, \quad \lambda' \in \Lambda, \tag{1.42}$$

for some constant $K = K(g, \sigma)$, and we call $\lambda(\cdot)$ an M-feedback control if $\lambda(\cdot)$ is Lipschitz-continuous, i.e.,

$$|\lambda(x) - \lambda(x')| \leqq M|x - x'| \quad \forall x, x' \in \mathcal{R}. \tag{1.43}$$

Consider the M-optimal cost

$$u(x, M) = \inf\{J_h(x, \lambda(\cdot)): \lambda(\cdot) M\text{-feedback control}\}, \tag{1.44}$$

for any $M > 0$, M destined to become infinite.

It is clear that $u(x, M) \geqq u(x)$ and, under reasonable assumptions we will have

$$u(x, M) \to u(x) \quad \forall x \in \mathcal{R} \quad \text{as } M \to \infty.$$

Moreover, sometimes the M-optimal cost is meaningful by itself.

THEOREM 1.4. *Let the assumptions of Theorem* 1.2 *and* (1.42) *hold. Then for any* M, $p > 0$ *there exist two constants* $C(M)$, $C > 0$ *depending only on* p, α, *the bound of* f, *and the constants of hypothesis* (1.1); $C(M)$ *depends also on* M *and the* $K(g, \sigma)$ *of* (1.42), *such that*

$$\begin{aligned} &u(x) - u_h(x) \leqq C[\sqrt{h} + \rho(r) + (r^{-1}\sqrt{h})^p] \quad \forall r > 0, \\ &u_h(x) - u(x, M) \leqq C(M)[\sqrt{h} + \rho(r) + (r^{-1}\sqrt{h})^p] \quad \forall r > 0, \end{aligned} \tag{1.45}$$

for any x *in* $\mathcal{R}$, h *in* $(0, 1]$ *and* $\rho(r)$ *given by* (1.31), *uniformly for* λ *in* Λ.

Proof. Starting from

$$u(x)-u_h(x)\leqq \sup\{J(x,\lambda_h)-J_h(x,\lambda(\cdot))\colon \lambda(\cdot)\},$$

$$e^{-\alpha t}\leqq \chi_h^{\alpha}(t)\leqq e^{-\alpha(t-h)}\quad \forall t\geqq 0$$

and in view of the estimate (1.18), we deduce the first part of (1.45) as in Theorem 1.3.

For the second part of (1.45) we use

$$u_h(x)-u(x,M)\leqq \sup\{J_h(x,\lambda(\cdot))-J(x,\lambda(\cdot))\colon \lambda(\cdot)\text{ any } M\text{-feedback control}\}$$

and we prove

$$E\sup\{|y_x(t,\lambda(\cdot))-y_x^h(t,\lambda(\cdot))|^p e^{-\alpha t}\colon t\geqq 0\}\leqq C(M)h^{p/2}, \tag{1.46}$$

as in Theorem 1.2, but now, $C(M)$ depends also on the Lipschitz constant M of the feedback control $\lambda(\cdot)$, as well as on the constant $K(g,\sigma)$ of (1.42). Thus, we complete the proof of the estimate (1.45). □

Until now, we have used only estimates on the stochastic state equation to obtain some bounds for the rate of convergence of the discrete HJB toward the continuous-time HJB.

Now we will look at the approximation problem in a more analytic way.

Suppose $\varphi(x)$ is a smooth function; then we can write

$$L_h(\lambda)\varphi=\frac{1}{2}\sigma^2\int_{-1}^{1}\varphi''(\cdot+gh+t\sigma\sqrt{h})(1-|t|)\,dt+g\int_0^1\varphi'(\cdot+tgh)\,dt \tag{1.47}$$

where the primes denote derivatives and we must take $\gamma=1$ in (1.4), i.e., for $g=g(x,\lambda)$, $\sigma=\sigma(x,\lambda)$,

$$L_h(\lambda)\varphi=\frac{1}{h}\left[\frac{1}{2}\varphi(\cdot+gh+\sigma\sqrt{h})+\frac{1}{2}\varphi\,\middle|\,(\cdot+gh-\sigma\sqrt{h})-\varphi(x)\right]. \tag{1.48}$$

First,

$$|L(\lambda)\varphi(x)-L_h\lambda\varphi(x)|\leqq C_\varphi h^{p/2}\quad \forall x\in\mathcal{R},\quad h\in(0,1], \tag{1.49}$$

and λ in Λ, and some constant C_φ depending on the bounds of g,σ,φ'', and the p-Hölder constant of φ'', i.e., the constant $K=K(\varphi'')$ satisfying

$$|\varphi''(x)-\varphi''(x')|\leqq K|x-x'|^p\quad \forall x,x'\in\mathcal{R}, \tag{1.50}$$

for some exponent $0<p\leqq 1$.

Let us define $[\varphi]_p$ as the infimum of the set of all constant C satisfying

$$\begin{aligned}&\inf\{L(\lambda)\varphi(y)\colon |y-x|\leqq C\sqrt{h}\}-Ch^{p/2}\leqq L_h(\lambda)\varphi(x)\\&\qquad\leqq \sup\{L(\lambda)\varphi(y)\colon |y-x|\leqq C\sqrt{h}\}+Ch^{p/2}\quad \forall h\in(0,1],\end{aligned} \tag{1.51}$$

for any x in Λ. It is clear that $|\varphi|_p$ can be bounded by the constant C_φ of (1.49). However, here we can do better:

(i) $[\varphi]_1$ is dominated by the bounds of the second derivative φ'' and the constants $C(g,\sigma)$, $K(g,\sigma)$ of hypothesis (1.1).

(ii) If $\sigma=\sigma(\lambda)$, i.e., constant in x, then $[\varphi]_p$ is dominated by the p-Hölder constant and the bound of the first derivative φ', and $C(g,\sigma)$, $K(g,\sigma)$.

(iii) If $g=0$ and $\sigma=\sigma(\lambda)$, then $[\varphi]_1$ is dominated by the bound of σ and does not depend on φ.

Suppose that f is bounded continuous and for some constant C, $K>0$, $0<p\leqq 1$,

$$\begin{aligned} &|f(x,\lambda)|\leqq C \quad \forall x\in\mathcal{R}, \quad \lambda\in\Lambda,\\ &|f(x,\lambda)-f(x',\lambda)|\leqq K|x-x'|^p \quad \forall x,x'\in\mathcal{R}, \quad \lambda\in\Lambda, \end{aligned} \tag{1.52}$$

and

$$\alpha>\alpha_p, \text{ the constant given by (1.20).} \tag{1.53}$$

THEOREM 1.5. *Under the assumptions* (1.1), (1.52), *and* (1.53) *there exists a constant* C *depending only on the constants* $C(g,\sigma)$, $K(g,\sigma)$, $C(f)$, $K(f)$ *of* (1.1), (1.52), *the constant* α *of* (1.53), *and the value* $[u]_p$ *with* u *being the maximum solution of the* HJB *equations* (1.37), *such that*

$$|u(x)-u_h(x)|\leqq Ch^{p/2} \quad \forall x\in\mathcal{R}, \quad h\in(0,1], \tag{1.54}$$

where $u_h(x)$ *is solution of the discrete* HJB *equation* (1.41) *with the finite difference operator* (1.48).

Proof. We remark that the fact that $[u]_p$ is finite is implicit. To check that the discrete HJB equation (1.41) has a solution, we rewrite it as follows:

$$\begin{aligned} &u_h=\inf\{\Pi_h^\alpha(\lambda)u_h+hf(\cdot,\lambda):\lambda\in\Lambda\} \quad \text{in } \mathcal{R},\\ &\Pi_h^\alpha(\lambda)\varphi=(1+\alpha h)^{-1}[hL_h(\lambda)\varphi-\varphi], \end{aligned} \tag{1.55}$$

and we note that the operator involved is a contraction map in the space of bounded continuous functions on $\mathcal{R}$.

First we will show that for some constants C, $K>0$ depending only on the constants in the assumptions (1.1), (1.52), and (1.53) such that

$$\begin{aligned} &|u_h(x)|+|u(x)|\leqq C \quad \forall x\in\mathcal{R},\\ &|u_h(x)-u_h(x')|+|u(x)-u(x')|\leqq K|x-x'|^p \quad \forall x,x'\in\mathcal{R}, \end{aligned} \tag{1.56}$$

for any h in $(0,1]$, $0<p\leqq 1$, the same p as in (1.52). It is relatively easy to obtain (1.50) for u from the stochastic representation (1.36); however, we prefer to use analytic arguments to present the technique used.

Consider the function

$$m(x,q,\varepsilon)=(\varepsilon^2+x^2)^{q/2} \quad \forall x\in\mathcal{R}, \tag{1.57}$$

for $q,\varepsilon>0$ fixed, and the solution $u(x)$ of the HJB equation (1.37). To prove the second part of (1.56) we look at the point (x_0,y_0) of $\mathcal{R}\times\mathcal{R}$ where the function

$$w(x,y)=u(x)-u(y)-Km(x-y,p,\varepsilon)m(x+y,q,1)$$

attains its maximum value, for a fixed K to be selected later. We want to show that $w(x_0,y_0)\leqq 0$ for an appropriate choice of K.

The extended operator

$$\begin{aligned} \tilde{L}(\lambda)\varphi(x,y)=&\tfrac{1}{2}\sigma^2(x,\lambda)\varphi''_{xx}+\sigma(x,\lambda)\sigma(y,\lambda)\varphi''_{xy}\\ &+\tfrac{1}{2}\sigma^2(y,\lambda)\varphi''_{yy}+g(x,\lambda)\varphi'_x+g(y,\lambda)\varphi'_y, \end{aligned} \tag{1.58}$$

is elliptic and satisfies

$$\tilde{L}(\lambda)[u(x)-u(y)]=L(\lambda)u(x)-L(\lambda)u(y).$$

After some calculations, we have

$$\begin{aligned}\tilde{L}(\lambda)[m(x-y,p,\varepsilon)m(x+y,q,1)]=&\frac{p}{2}[\sigma(x,\lambda)-\sigma(y,\lambda)]^2\\&\cdot m(x-y,p-2,\varepsilon)m(x+y,q,1)\\&\cdot[(p-1)(x-y)^2m(x-y,-2,\varepsilon)+1]\\&+\frac{q}{2}[\sigma(x,\lambda)+\sigma(y,\lambda)]^2m(x-y,p,\varepsilon)\\&\cdot m(x+y,1,q-2)\\&\cdot[(q-1)(x+y)^2m(x+y,-2,1)]\\&+pq[\sigma(x,\lambda)+\sigma(y,\lambda)][\sigma(x,\lambda)-\sigma(y,\lambda)]\\&\cdot(x-y)(x+y)m(x-y,p-2,\varepsilon)\\&\cdot m(x+y,q-2,1)+p[g(x,\lambda)-g(y,\lambda)]\\&\cdot(x-y)m(x-y,p-2,\varepsilon)m(x+y,q,1)\\&+q[g(x,\lambda)+g(y,\lambda)]\\&\cdot(x+y)m(x-y,p,\varepsilon)m(x+y,q-2,1),\end{aligned}$$

which shows that

$$\tilde{L}(\lambda)[m(x-y,p,\varepsilon)m(x+y,q,1)]\leqq(\alpha_p-qC)m(x-y,p,\varepsilon)m(x+y,q,1), \tag{1.59}$$

where α_p is the constant defined by (1.20) and C is a constant independent of λ, x, y, ε, p, and $0<q<1$. We choose $q>0$ such that $\alpha-\alpha_p+qC\geqq\alpha_0>0$.

Now, by means of the maximum principle, we have $L(\lambda)w(x_0,y_0)\leqq 0$, i.e.,

$$L(\lambda)u(x_0)-L(\lambda)u(y_0)\leqq(\alpha-\alpha_0)Km(x_0-y_0,p,\varepsilon)m(x_0+y_0,1,q), \tag{1.60}$$

assuming that u is smooth and after using (1.59). But, from HJB equation (1.37) we deduce

$$\alpha[u(x_0)-u(y_0)]\leqq[K(f)+(\alpha-\alpha_0)K]m(x_0-y_0,p,\varepsilon)m(x_0+y_0,1,q),$$

where $K(f)$ is the p-Hölder Lipschitz of f in (1.52). Hence, if we choose $K=\alpha_0^{-1}K(f)$, then we conclude that $w(x_0,y_0)\leqq 0$. Therefore, we should have

$$u(x)-u(y)\leqq Km(x-y,p,\varepsilon)m(x+y,q,1).$$

Because the constant K does not depend on ε, q, we send ε, q to zero to obtain the second part of (1.56) for u, assuming that u is smooth.

Similarly, we show the Hölder-continuous estimate for u_h. In that case we use the extended operator

$$\begin{aligned}&\tilde{L}_h(\lambda)\varphi(x,y)=\frac{1}{h}\left[\frac{1}{2}\varphi(z^+(x,\lambda),z^+(y,\lambda)+\frac{1}{2}\varphi(z^-(x,\lambda),z^-(y,\lambda))-\varphi(x,y)\right],\\&z^{\pm}(\cdot,\lambda)=\cdot+g(\cdot,\lambda)h\pm\sigma(\cdot,\lambda)\sqrt{h}.\end{aligned} \tag{1.61}$$

Note that if u is not smooth then we have to approximate u by a smooth function u_e, either by regularization, i.e., $\sigma+\varepsilon$ replaces σ, or by the so-called infimum convolution. The proof of the first part of (1.56) uses a technique analogous to the above.

Let us prove the estimate (1.54). Consider the function

$$w(x,y)=u_h(x)-u(y)-C_1m(x-y,p,\varepsilon)m(x+y,q,1)-C_2h^{p/2}$$

for some constants C_1, C_2, q, $\varepsilon > 0$ to be selected later, and let (x_0, y_0) be a point where $w(x, y)$ attains its maximum value. A calculation similar to the one to obtain (1.59) shows that

$$\tilde{L}_h(\lambda)[m(x-y, p, \varepsilon)m(x+y, q, 1)] \leqq (\alpha_p - rq)m(x-y, p, \varepsilon)m(x+y, q, 1), \tag{1.62}$$

for any x, y in $\mathcal{R}$, λ in Λ, h, q in $(0, 1]$, some constant $r > 0$ and the same α_p of (1.20). We take $q > 0$ such that $\alpha_p - rq \leqq \alpha - \alpha_0$, $\alpha_0 > 0$.

Because $\tilde{L}_h(\lambda)w(x_0, y_0) \leqq 0$ we deduce

$$L_h(\lambda)u_h(x_0) - L_h(\lambda)u(y_0) \leqq (\alpha - \alpha_0)C_1 m(x_0 - y_0, p, \varepsilon)m(x_0 + y_0, q, 1),$$

and in view of (1.51),

$$L_h(\lambda)u(y_0) \leqq L(\lambda)u(y_1) + [u]_p h^{p/2}, \qquad |y_0 - y_1| \leqq [u]_p \sqrt{h}. \tag{1.63}$$

From the HJB equations satisfied by u_h and u we obtain

$$\alpha[u_h(x_0) - u(y_1)] \leqq \sup\{|f(x_0, \lambda) - f(y_1, \lambda)| : \lambda \in \Lambda\} + (\alpha - \alpha_0)Cm(x_0 - y_0, p, \varepsilon)m(x_0 + y_0, q, 1) + [u]_p h^{p/2},$$

and by means of (1.52), (1.56), (1.63) we get

$$|f(x_0, \lambda) - f(y_1, \lambda)| + |u(y_0) - u(y_1)| \leqq [K(f) + K(u)]m(x_0 - y_0, p, \varepsilon),$$

provided $\varepsilon = [u]_p \sqrt{h}$.

Collecting all, we have

$$\alpha[u_h(x_0) - u(y_0)] \leqq [K(f) + K(u) + (\alpha - \alpha_0)C_1]m(x_0 - y_0, p, \varepsilon) \cdot m(x_0 + y_0, q, 1) + [u]_p h^{p/2}.$$

Hence, if we choose

$$C_1 = \alpha_0^{-1}[K(f) + K(u)], \quad C_2 = [u]_p, \quad \varepsilon = [u]_p \sqrt{h},$$

then $w(x_0, y_0) \leqq 0$, i.e.,

$$u_h(x) - u(y) \leqq C_1 m(x - y, p, \varepsilon)m(x + y, q, 1) + C_2 h^{p/2}, \tag{1.64}$$

for any x, y in $\mathcal{R}$, h, q in $(0, 1]$. Letting q vanish and taking $x = y$, we establish one side of (1.54).

Reversing the role of u_h and u we complete the proof. □

Remark 1.5. Note that in the proof of Theorem 1.5 we assume implicitly that the function u is smooth. However, once the estimates have been established, we can remove that assumption on u, only $[u]_p$ needs to be finite.

1.4. Extension and comments. The fact that the functions g, σ are bounded is not really important, we need only to assume linear growth, i.e.,

$$|g(x)| + |\sigma(x)| \leqq C(1 + |x|) \quad \forall x \in \mathcal{R}, \tag{1.65}$$

for some constant $C = C(g, \sigma)$. In this case the estimate (1.18) of Theorem 1.2 becomes

$$E \sup\{|y_x(t, \lambda^h) - y_x^h(t, \lambda(\cdot))|^p e^{-\alpha t} : t \geqq 0\} \leqq C(1 + |x|^2)^{p/2} h^{p/2} \quad \forall x \in \mathcal{R}, \tag{1.66}$$

for some constants C, $\alpha > 0$.

To adapt Theorem 1.1 to the time-dependent case, we modify the construction (1.6), (1.7), for instance,

$$\tau(x, t, \lambda, h, w) = \inf\{s \geqq 0 : g(x, s + t, \lambda)(s - h) + \sigma(x, s + t, \lambda)w(s) \text{ equals } \pm\delta(x, t, \lambda, h)\}. \tag{1.67}$$

A generalization to dimension $d \geqq 2$ is possible but more delicate. Let us describe the procedure. We write σ as the matrix formed by the column vectors $\sigma_1, \sigma_2, \cdots, \sigma_n$; the drift vector g is expressed as $g = g^1 e_1 + \cdots + g^n e_n$, where g^i are scalars and e_i are vectors in the direction σ_i, i.e., $\sigma_i = \sigma^i e_i$, σ^i is scalar. We want to define τ_i as the first time for which

$$g^i e_i(\tau_i - h) + \sigma^i e_i w_i(\tau_i) = \pm \sigma^i \gamma^i \sqrt{h}\, e_i.$$

This is the same as cancelling the vector e_i and defining τ_i as in (1.6) with g, σ, γ, w, replaced by $g^i, \sigma^i, \gamma^i, w_i$. Then we are interested in the stopping time $\tau = \min\{\tau_i: i = 1, \cdots, n\}$, which is the first exit time of the box in $\mathcal{R}^n$ bounded by the hyperplane $z_i = \pm \sigma^i \gamma^i \sqrt{h}$, $z = (z_1, \cdots, z_n)$ in $\mathcal{R}^n$, for a Wiener process in $\mathcal{R}^n$ with drift $(g^1, \cdots, g^n)$ and diffusion term the diagonal matrix $(\sigma^1, \cdots, \sigma^n)$, starting at the point $(-g^1\sqrt{h}, \cdots, -g^n\sqrt{h})$. Details of this construction will be presented in a future work.

In Theorems 1.3, 1.4, and 1.5 we can allow the functions g, σ to satisfy (1.65) and the function f to have polynomial growth, i.e.,

$$\begin{aligned} &|f(x, \lambda)| \leqq C(1 + x^2)^{q/2} \quad \forall x \in \mathcal{R}, \\ &|f(x, \lambda) - f(y, \lambda)| \leqq K|x - y|^p (1 + x^2 + y^2)^{r/2} \quad \forall x, y \in \mathcal{R}, \end{aligned} \tag{1.68}$$

for $q > 0$; $0 < q \leqq 1$, $r = \max\{q - p, 0\}$. The estimate (1.54) is modified accordingly. For the estimates (1.32) and (1.45) we use

$$\rho(r) = \inf\{|f(x, \lambda) - f(y, \lambda)|(1 + x^2 + y^2)^{q/2}: x, y \text{ in } \mathcal{R}, \lambda \text{ in } \Lambda\}. \tag{1.69}$$

A discretization in Λ can also be incorporated. In that case, a term of the form

$$r(h) = \sup\{\inf\{|l(x, \lambda) - l(x, \lambda')|: \lambda' \in \Lambda(h)\}: x \in \mathcal{R}^d, \lambda \in \Lambda, l = f, g, \sigma\} \tag{1.70}$$

will appear in the estimates (1.32), (1.45), and (1.54) of Theorems 1.3, 1.4, and 1.5. Here $\Lambda(h)$ is a discretization of Λ.

The constant $\alpha > 0$ can be replaced by a function $\alpha(x, \lambda)$.

The fact that we made only the discretization in the time variable is just the first step. To discretize the space variable, we can add the second part of condition (0.7), as in the next section. An alternative is to use finite elements to solve the discrete HJB of the type (1.41). This issue is reserved for a future work.

2. General problems. In this section we will consider the typical control problem (0.1)-(0.5) in a bounded open subset D of $\mathcal{R}^d$.

Let g and σ be bounded continuous functions from $\mathcal{R}^d \times \Lambda$ into $\mathcal{R}^d$ and $\mathcal{R}^d \times \mathcal{R}^n$, respectively, such that $g = (g_i, i = 1, \cdots, d)$, $\sigma = (\sigma_{ik}, i = 1, \cdots, d, k = 1, \cdots, n)$,

$$\begin{aligned} &|g(x, \lambda)|\sigma(x, \lambda)| \leqq C \quad \forall x \in \mathcal{R}^d \quad \forall \lambda \in \Lambda, \\ &|g(x, \lambda) - g(x', \lambda)| + |\sigma(x, \lambda) - \sigma(x', \lambda)| \leqq K|x - x'| \quad \forall x, x' \in \mathcal{R}^d, \quad \lambda \in \Lambda, \end{aligned} \tag{2.1}$$

for some constants $C = C(g, \sigma)$, $K = K(g, \sigma)$, some locally compact metric space Λ and where $|\cdot|$ denotes the Euclidean norm in the corresponding space.

On a complete Wiener space $(\Omega, P, \mathcal{F}, \mathcal{F}(t), w(t), t \geqq 0)$ in $\mathcal{R}^n$ we consider the state equation

$$dy(t) = g(y(t), \lambda(t))\, dt + \sigma(y(t), \lambda(t))\, dw(t), \quad t > 0, \quad y(0) = x, \tag{2.2}$$

where the control $(\lambda(t), t \geqq 0)$ is a progressively measurable process taking values in Λ. Denote by τ the first exit time of $\bar{D}$, closure of D, for the process $(y(t), t \geqq 0)$, i.e.,

$$\tau = \inf\{t \geqq 0: y(t) \notin \bar{D}\}. \tag{2.3}$$

For a given real bounded-continuous function f on $\mathcal{R}^d \times \Lambda$ such that

$$
\begin{aligned}
&|f(x,\lambda)| \leqq C \quad \forall x \in \mathcal{R}^d, \quad \lambda \in \Lambda, \\
&|f(x,\lambda) - f(x',\lambda)| \leqq K|x-x'|^p \quad \forall x, x' \in \mathcal{R}^d, \quad \lambda \in \Lambda,
\end{aligned} \tag{2.4}
$$

for some constants C, $K > 0$, $0 < p \leqq 1$, we define

$$
J(x,\lambda) = E\left\{ \int_0^\tau f(y(t), \lambda(t))\, e^{-\alpha t}\, dt \right\}, \qquad \alpha > 0, \tag{2.5}
$$

and the optimal cost function

$$
u(x) = \inf\{J(x,\lambda): \text{ any control process } \lambda\}. \tag{2.6}
$$

The HJB equation is

$$
\alpha u = \inf\{L(\lambda)u + f(\cdot,\lambda): \lambda \in \Lambda\} \quad \text{in } D, \quad u = 0 \quad \text{on } \partial D, \tag{2.7}
$$

where the differential operator

$$
L(\lambda) = \frac{1}{2} \sum_{i,j=1}^{d} \left(\sum_{k=1}^{n} \sigma_{ik}(\cdot,\lambda)\sigma_{jk}(\cdot,\lambda) \right) \partial_{ij} + \sum_{i=1}^{d} g_i(\cdot,\lambda)\partial_i, \tag{2.8}
$$

and the bounded domain D has a uniform exterior sphere, i.e.,

$$
\begin{aligned}
&\text{there exists } r > 0 \text{ such that for any } x \text{ in } \partial D \text{ there is } y \text{ in } \mathcal{R}^d \setminus D \text{ such that} \\
&\{z: |y-z| \leqq r\} \cap \bar{D} = \{x\},
\end{aligned} \tag{2.9}
$$

and $L(\lambda)$ is not degenerate on the boundary, i.e.,

$$
\sum_{k=1}^{n} \sum_{i=1}^{d} |\sigma_{ik}(x,\lambda)\eta_i(x)| \geqq \nu_0 > 0 \quad \forall x \in \partial D, \quad \lambda \in \Lambda, \tag{2.10}
$$

with $\eta = (\eta_1, \cdots, \eta_d)$ being a normal direction to ∂D.

In § 2.1 we will give some properties of the finite difference operator (0.6). Next, we study the discrete HJB equations and its associated Markov chain. We present the main estimate in § 2.3 and then we give some comments and extensions.

2.1. The finite difference operator. Recall the operator (0.6),

$$
\begin{aligned}
L_h(\lambda)\varphi(x) = h^{-1} \sum_{k=1}^{n} \{ &\beta_k^+(x,\lambda,h)[\varphi(x+\gamma_k^+(x,\lambda,h)) - \varphi(x)] \\
&+ \beta_k^-(x,\lambda,h)[\varphi(x+\gamma_k^-(x,\lambda,h)) - \varphi(x)]\},
\end{aligned} \tag{2.11}
$$

where $\beta_k^\pm$, $\gamma_k^\pm$ are bounded Borel-measurable functions in x, λ for h fixed,

$$
\beta_k^\pm(x,\lambda,h) \geqq 0, \quad x + \gamma_k^\pm(x,\lambda,h) \in \mathcal{R}_h^d \quad \forall x \in \mathcal{R}_h^d, \quad \lambda \in \Lambda. \tag{2.12}
$$

The h-finite difference grid $\mathcal{R}_h^d$ is given, $0 < h \leqq 1$.

We denote

$$
L_h(k,\lambda)\varphi(x) = h^{-1}\{\beta_k^+[\varphi(x+\gamma_k^+) - \varphi(x)] + \beta_k^-[\varphi(x+\gamma_k^-) - \varphi(x)]\}, \tag{2.13}
$$

and

$$
\gamma_k^0 = (\beta_k^+ + \beta_k^-)^{-1}(\beta_k^+\gamma_k^+ + \beta_k^-\gamma_k^-), \tag{2.14}
$$

where the variables x, λ, h have been omitted.

We can rewrite

$$L_h(k,\lambda)\varphi(x) = h^{-1}\{\beta_k^+[\varphi(x+\gamma_k^+)-\varphi(x+\gamma_k^0)] + \beta_k^-[\varphi(x+\gamma_k^-)-\varphi(x+\gamma_k^0)]+(\beta_k^+ + \beta_k^-)[\varphi(x+\gamma_k^0)-\varphi(x)]\},$$

and when φ is smooth,

$$\varphi(x+\gamma_k^\pm)-\varphi(x+\gamma_k^0) = \sum_{i=1}^d (\gamma_{ik}^\pm - \gamma_{ik}^0)\int_0^1 \partial_i\varphi(x+\gamma_k^0+t(\gamma_k^\pm-\gamma_k^0))\,dt,$$

$$\varphi(x+\gamma_k^0)-\varphi(x) = \sum_{i=1}^d \gamma_{ik}^0 \int_0^1 \partial_i\varphi(x+t\gamma_k^0)\,dt,$$

with $\gamma_{ik}^\pm$, γ_{ik}^0 being the components of $\gamma_k^\pm$, γ_k^0. Using the fact that

$$\gamma_k^\pm - \gamma_k^0 = \pm\beta_k^\pm(\beta_k^+ + \beta_k^-)^{-1}(\gamma_k^+ - \gamma_k^-),$$

we have

$$\beta_k^+[\varphi(x+\gamma_k^+)-\varphi(x+\gamma_k^0)]+\beta_k^-[\varphi(x+\gamma_k^-)-\varphi(x+\gamma_k^0)]$$
$$= \beta_k^+\beta_k^-(\beta_k^+ + \beta_k^-)^{-2}\sum_{i,j=1}^d (\gamma_{ik}^+ - \gamma_{ik}^-)(\gamma_{jk}^+ - \gamma_{jk}^-)$$
$$\cdot \int_0^1 dt \int_{t\beta_k^+}^{t\beta_k^-} \partial_{ij}\varphi(x+\delta_k(s))\,ds,$$

where

$$\delta_k(s) = (\beta_k^+ + \beta_k^-)^{-1}[(\beta_k^+ - s)\gamma_k^+ + (\beta_k^- - s)]\gamma_k^-.$$

If

$$\chi_k(s) = \begin{cases} 2(\beta_k^+ + \beta_k^-)^{-1}[1 - s(\beta_k^-)^{-1}] & \text{if } s \geqq 0, \\ 2(\beta_k^+ + \beta_k^-)^{-1}[1 + s(\beta_k^+)^{-1}] & \text{if } s \leqq 0, \end{cases}$$

then

$$hL_h(k,\lambda)\varphi(x) = \frac{1}{2}\beta_k^+\beta_k^-(\beta_k^+ + \beta_k^-)^{-1}\sum_{i,j=1}^d (\gamma_{ik}^+ - \gamma_{ik}^-)(\gamma_{jk}^+ - \gamma_{jk}^-)$$
$$\cdot \int_{-\beta_k^+}^{\beta_k^-} \partial_{ij}\varphi(x+\delta_k(s))\chi_k(s)\,ds$$
$$+ \sum_{i=1}^d (\beta_k^+\gamma_{ik}^+ + \beta_k^-\gamma_{ik}^-)\int_0^1 \partial_i\varphi(x+t\gamma_k^0)\,dt. \tag{2.15}$$

Note that

$$\int_{-\beta_k^+}^{\beta_k^-} \chi_k(s)\,ds = 1$$

and that $\delta_k(s)$, γ_k^0 are convex combinations of γ_k^+ and γ_k^-.

Therefore, let us assume that for some constant $C > 0$ and any $i, j = 1, \cdots, d$, we have

$$\left|\sum_{k=1}^n [\sigma_{ik}\sigma_{jk}h - (\beta_k^+ + \beta_k^-)^{-1}\beta_k^+\beta_k^-(\gamma_{ik}^+ - \gamma_{ik}^-)(\gamma_{jk}^+ - \gamma_{jk}^-)]\right| \leqq C^2h^{3/2},$$

$$\left|g_ih - \sum_{k=1}^n (\beta_k^+\gamma_{ik}^+ + \beta_k^-\gamma_{ik}^-)\right| \leqq Ch^{3/2}, \tag{2.16}$$

$$|\gamma_{ik}^+| + |\gamma_{ik}^-| \leqq Ch^{1/2}$$

uniformly for the variables x in $\mathcal{R}_h^d$, λ in Λ, h in $(0, 1]$. Then we have the following estimate:

$$|L_h(\lambda)\varphi(x) - L(\lambda)\varphi(x)| \leqq C_\varphi h^{p/2} \quad \forall x, \lambda, h, \tag{2.17}$$

where C_φ is a constant depending only on the bounds of functions $g, \sigma, \partial_{ij}, \varphi$ and the Hölder-continuous constants of the second derivatives $\partial_{ij}\varphi$ with exponent $0 < p \leqq 1$, i.e., the constant $K(\partial_{ij}\varphi)$ of (2.4) for $\partial_{ij}\varphi$ in lieu of f.

Typical examples where the assumption (2.16) holds are the following cases: any $\gamma_{ik}^{\pm}$, $\beta_k^{\pm}$, g_{ik}, σ_{ik} satisfying

$$\begin{aligned}
&\gamma_{ik}^{\pm} = g_{ik}(x, \lambda, h)\beta_k^{-2}(x, \lambda, h)h \pm \sigma_{ik}(x, \lambda, h)\beta_k^{-1}(x, \lambda, h)\sqrt{h},\\
&\beta_k^{\pm} = \frac{1}{2}\beta_k(x, \lambda, h), \quad \beta_k(x, \lambda, h) > 0, \quad k = 1, \cdots, n,\\
&\left| g_i(x, \lambda) - \sum_{k=1}^{n} g_{ik}(x, h, \lambda) \right| \leqq Ch^{1/2},\\
&|\sigma_{ik}(x, \lambda)\sigma_{jk}(x, \lambda) - \sigma_{ik}(x, \lambda, h)\sigma_{jk}(x, \lambda, h)| \leqq Ch^{1/2},
\end{aligned} \tag{2.18}$$

uniformly in x, λ, h and for some constant $C > 0$. A more classic possibility is to choose $n = d$,

$$\gamma_{ik}^{\pm}(x, \lambda, h) = \begin{cases} \pm\sqrt{h} & \text{if } i = k,\\ 0 & \text{otherwise,} \end{cases} \tag{2.19}$$

and accordingly the coefficients $\beta_k^{\pm}(x, \lambda, h)$ to insure (2.16). Also, we may take $n = d + 1$,

$$\begin{aligned}
&\gamma_{ik}^{\pm} = \pm\sigma_{ik}(x, \lambda, h)\beta_k^{-1}(x, \lambda, h)\sqrt{h} \quad \text{if } k = 1, \cdots, n-1,\\
&\beta_k^{\pm} = \tfrac{1}{2}\beta_k(x, \lambda, h) > 0 \quad \text{if } k = 1, \cdots, n-1,\\
&\gamma_{in}^{\pm} = g_i(x, \lambda, h)\beta_n^{-2}(x, \lambda, h)h, \qquad \beta_n > 0,\\
&\beta_n^{\pm} = \beta_n(x, \lambda, h), \qquad \beta_n^{-} = 0, \quad \sigma_{in}(x, \lambda) = 0 \quad \forall\\
&|g_i(x, \lambda) - g_i(x, \lambda, h)| \leqq Ch^{1/2} \quad \forall i,\\
&|\sigma_{ik}(x, \lambda)\sigma_{jk}(x, \lambda) - \sigma_{ik}(x, \lambda, h)\sigma_{jk}(x, \lambda, h)| \leqq Ch^{1/2} \quad \forall i, jk, \quad k \neq n,
\end{aligned} \tag{2.20}$$

uniformly in x, λ, h for some constant $C > 0$.

When the differential operator (2.8) is degenerate with constant order of degeneration, i.e.,

$$L(\lambda) = \frac{1}{2}\sum_{i,j=1}^{m}\left(\sum_{k=1}^{n}\sigma_{ik}(\cdot, \lambda)\sigma_{jk}(\cdot, \lambda)\right)\partial_{ij} + \sum_{i=1}^{d} g_i(\cdot, \lambda)\partial_i, \tag{2.21}$$

where $0 \leqq m \leqq d$, $n \geqq 0$ and clearly $d - m$ is the order of degeneration, it is convenient to choose (2.18) or (2.20) instead of (2.19). In this case the constant C_φ of (2.17) will depend only on the constants $K(\partial_{ij}, \varphi)$ and bounds of $g, \sigma, \partial_{ij}\varphi$, for $i, j = 1, \cdots, m$.

Denote by $\mathcal{R}_h^d$ an h-finite difference grid in $\mathcal{R}^d$, i.e.,

$$\begin{aligned}
&\mathcal{R}_h^d = \{x = (x_1, \cdots, x_d) : \forall i = 1, \cdots, d, \exists k = 0, \pm 1, \cdots, \text{such that } x_i = hr_i(k)\},\\
&1 \leqq r_i(k) - r_i(k-1) < 2 \quad \forall k = 0, \pm 1, \cdots.
\end{aligned} \tag{2.22}$$

For the open bounded subset D of $\mathcal{R}^d$ we denote

$$D_h = \{x \in D \cap \mathcal{R}_h^d : x + \gamma_k^{\pm}(x, \lambda, h) \in \bar{D}, \forall \lambda, k\}, \tag{2.23}$$

and its boundary

$$\begin{aligned}&\Gamma_h = \cup\{\Gamma_k^+(\lambda)\cup\Gamma_k^-(\lambda)\colon \lambda, k\},\\ &\Gamma_k^\pm(\lambda) = \{x\in D\cap R_h^d\colon x+\gamma_k^\pm(x,\lambda,h)\in \bar{D}\cap \mathscr{R}_h^d \text{ and } x+\gamma_k^\mp(x,\lambda,h)\notin \bar{D}\cap\mathscr{R}_h^d\},\end{aligned} \tag{2.24}$$

for a fixed h in $(0, 1]$.

Under the assumptions (2.11), (2.12) we can easily prove the discrete maximum principle for the finite difference operator $L_h(\lambda)$. It is as follows. If a function $u_h(x)$ defined on $\bar{D}_h = D_h\cup\Gamma_h$ attains its maximum value at some point x_0, then

$$\begin{aligned}&\text{(i)}\quad \text{If } x_0\in D_h, \text{ then } L_h(\lambda)u_h(x_0)\leqq 0 \quad \forall\lambda\in\Lambda;\\ &\text{(ii)}\quad \text{If } x_0\in\Gamma_k^\pm(\lambda), \text{ then } \nabla_k^\pm(\lambda)u_h(x_0)\leqq 0,\end{aligned} \tag{2.25}$$

where $\nabla_k^\pm(\lambda)$ is the operator given by

$$\nabla_k^\pm(\lambda)\varphi(x) = h^{-1}\beta_k^\pm(x,\lambda,h)[\varphi(x+\gamma_k^\pm(x,\lambda,h))-\varphi(x)], \tag{2.26}$$

for any φ.

2.2. Study of the discrete equation. Here, we are interested in the discrete HJB equation (0.8), i.e., in finding a function $u_h(x)$, x in $\bar{D}_h$ such that

$$\alpha u_h = \inf\{L_h(\lambda)u_h + f(\cdot,\lambda)\colon \lambda\in\Lambda\} \quad \text{in } D_h, \quad u_h = 0 \quad \text{on } \Gamma_h, \tag{2.27}$$

where D_h, Γ_h, $\bar{D}_h$ are defined by (2.23), (2.24), and the finite difference operator $L_h(\lambda)$ is given by (2.11), (2.12).

First, we will associate an optimal control problem of a Markov chain to the HJB equation (2.27).

Let

$$G(\xi) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\xi}\exp\left(-\frac{t^2}{2}\right)dt, \qquad -\infty\leqq\xi\leqq+\infty, \tag{2.28}$$

and

$$\begin{aligned}&G(\tilde{\beta}_1^+) = \beta^{-1}\beta_1^+, \qquad G(\tilde{\beta}_1^-) = \beta^{-1}(\beta_1^+ + \beta_1^-),\\ &G(\tilde{\beta}_k^+) = \beta^{-1}(\beta_1^+ + \beta_1^- + \cdots + \beta_k^+), \qquad k = 2,\cdots,n,\\ &G(\tilde{\beta}_k^-) = \beta^{-1}(\beta_1^+ + \beta_1^- + \cdots + \beta_k^+ + \beta_k^-), \qquad k = 2,\cdots,n-1,\\ &\beta = \beta_1^+ + \beta_1^- + \cdots + \beta_n^+ + \beta_n^-, \quad \beta_0^- = -\infty, \quad \tilde{\beta}_n^- = +\infty\end{aligned} \tag{2.29}$$

where the variables x, λ, h have been omitted. For a random variable η with Gaussian density (2.28), we define the random fields $\xi_k^\pm(w) = \xi_k^\pm(x,\lambda,h;\eta)$ by

$$\xi_k^\pm(w) = \begin{cases}1 & \text{if } w\in A_k^\pm,\\ 0 & \text{otherwise}\end{cases} \tag{2.30}$$

where

$$\begin{aligned}&A_k^+ = \{w\colon \tilde{\beta}_{k-1}^- < \eta(w) < \tilde{\beta}_k^+\}, \qquad \tilde{\beta}_0^- = -\infty,\\ &A_k^- = \{w\colon \tilde{\beta}_k^+ < \eta(w) < \tilde{\beta}_k^-\}, \qquad k = 1, 2, \cdots, n.\end{aligned} \tag{2.31}$$

Suppose we are given a sequence $(\eta_i: i=1,2,\cdots)$ of independent random variables with the same Gaussian density (2.28) in a complete probability space $(\Omega, P, \mathcal{F})$. Then we consider the following controlled Markov chain:

$$
\begin{aligned}
X^{i+1} = X^i + \sum_{k=1}^{n} [&\gamma_k^+(X^i, \lambda(X^i), h)\xi_k^+(X^i, \lambda(X^i), h; \eta_{i+1}) \\
&+ \gamma_k^-(X^i, \lambda(X^i), h)\xi_k^-(X^i, \lambda(X^i), h; \eta_{i+1})], \qquad i=0,1,\cdots, \\
&X^0 \text{ given in } \mathcal{R}_h^d,
\end{aligned} \tag{2.32}
$$

where $\lambda = \lambda(\cdot)$ is a feedback control, i.e., a Borel-measurable function from $\mathcal{R}^d$ into Λ; actually it suffices that λ be defined only on $\mathcal{R}_h^d$.

A simple calculation shows that the transition probability operators of the Markov chain (2.32) is given by

$$
\begin{aligned}
&E\{\varphi(X^{i+1}) \mid X^i = x\} = \pi_h(\lambda(x))\varphi(x), \\
&\pi_h(\lambda)\varphi(x) = \beta^{-1}(x, \lambda, h) \sum_{k=1}^{n} [\beta_k^+(x, \lambda, h)\varphi(x + \gamma_k^+(x, \lambda, h)) \\
&\qquad + \beta_k^-(x, \lambda, h)\varphi(x + \gamma_k^-(x, \lambda, h))] \quad \forall \varphi, \\
&\beta(x, \lambda, h) = \sum_{k=1}^{n} [\beta_k^+(x, \lambda, h) + \beta_k^-(x, \lambda, h)] \quad \text{for any } x, \lambda, h.
\end{aligned} \tag{2.33}
$$

Standard arguments of the discrete optimal control theory (e.g., Bertsekas and Shreve [3], Gihkman and Skorokhod [10], Ross [32]) show that the optimal cost function

$$
\begin{aligned}
&u_h(x) = \inf\{J_h(x, \lambda): \lambda(\cdot) \text{ any feedback control}\}, \\
&J_h(x, \lambda) = E\left\{\sum_{i=0}^{\nu-1} hf(X^i, \lambda(X^i))[q_h(X^i, \lambda(X^i))]^i \mid X^0 = x\right\}, \\
&q_h(x, \lambda) = [h\alpha + \beta(x, \lambda, h)]^{-1}\beta(x, \lambda, h), \\
&\nu = \inf\{i=0,1,\cdots: X^i \in \Gamma_h\}
\end{aligned} \tag{2.34}
$$

satisfies the discrete HJB equation (2.27). Note that

$$
\begin{aligned}
&E\{X^{i+1} \mid X^i = x\} = x + \sum_{k=1}^{n} \gamma_k^0, \\
&\text{Var}\{X^{i+1} \mid X^i = x\} = \sum_{k,i,j} [(\gamma_k^+ - \gamma_i^0)(\gamma_k^+ - \gamma_i^0)^* + (\gamma_k^- - \gamma_i^0)(\gamma_k^- - \gamma_i^0)^*], \\
&\gamma_k^0 = \beta^{-1}(\beta_k^+\gamma_k^+ + \beta_k^-\gamma_k^-),
\end{aligned} \tag{2.35}
$$

which are related to the condition (2.16).

We remark that the random fields (2.30) possess the following property:

$$
\begin{aligned}
P\{w \in \Omega: \xi_k^{\pm}(x, \lambda, h; \eta(w)) \neq \xi_k^{\pm}(x', \lambda', h; \eta(w))\} &\leqq 2\beta^{-1}(x, \lambda, h) \\
&\cdot \sum_{k=1}^{n} [|\beta_k^+(x, \lambda, h) - \beta_k^+(x', \lambda', h)| + |\beta_k^-(x, \lambda, h) - \beta_k^-(x', \lambda', h)|],
\end{aligned} \tag{2.36}
$$

for any x, λ, h. All the above properties are useful for directly studying the dependence on the data of the optimal cost (2.34) (cf. [24]).

On the other hand, we can use the technique of barrier functions used in continuous time control problems (e.g., Lions [20], Lions and Menaldi [21]) to construct subsolutions of the discrete HJB equation; this method uses the assumptions (2.9) and (2.10). So, we suppose that:

(2.37) There exist functions $\bar{u}_h(x)$ defined on $\mathcal{R}^d$ that are bounded and Hölder-continuous uniformly in h with exponent $0<p\leqq 1$ such that for some constants $\beta_p \geqq 0$, $K>0$ we have $L_h(\lambda)\bar{u}_h \leqq -1+\beta_p \bar{u}_h$, in D_h, $\forall \lambda \in \Lambda$, $\bar{u}_h(x)=0$, $\forall x \in \Gamma_h$, $h\in(0,1]$, $|\bar{u}_h(x)-\bar{u}_h(x')| \leqq K|x-x'|^p$, $\forall x, x' \in \mathcal{R}^d$.

THEOREM 2.1. *Let us assume* (2.11), (2.12), (2.37). *Then for any bounded Borel-measurable function* $f(x,\lambda)$ *and any constant* $\alpha>0$ *there exist a unique solution of the discrete* HJB *equation* (2.27). *Moreover, for two data* $f, \tilde{f}$ we have

$$\|u_h-\tilde{u}_h\| \leqq \alpha^{-1}\|f-\tilde{f}\| \quad \forall h\in(0,1], \tag{2.38}$$

where $\tilde{u}_h$ *denotes the solution corresponding to* $\tilde{f}$ *and* $\|\cdot\|$ *is the supremum norm. Furthermore, if* $\alpha \geqq \beta_p$ *in* (2.37) *then*

$$|u_h(x)| \leqq \|f\|\bar{u}_h(x) \quad \forall x \in \bar{D}_h, \quad h\in(0,1]. \tag{2.39}$$

Proof. It is possible to use the Markov chain (2.32) to establish the results as in [24], but we prefer to illustrate its analytic counterpart.

First of all, we rewrite the discrete HJB equation (2.27) as

$$u_h = \inf\{g_h(\cdot,\lambda)\pi_h(\lambda)u_h + f_h(\cdot,\lambda) : \lambda\in\Lambda\} \quad \text{in } D_h, \quad u_h=0 \quad \text{on } \Gamma_h, \tag{2.40}$$

where

$$f_h(x,\lambda) = h[h\alpha+\beta(x,\lambda,h)]^{-1}f(x,\lambda), \tag{2.41}$$

and $\pi_h(\lambda)$, $\beta(x,\lambda,h)$, $q_h(x,\lambda)$ are defined by (2.33), (2.34). If we denote by $T_h(u_h)$ the right-hand side of (2.40), then T_h is a contraction mapping on the space of bounded Borel-measurable functions from $\bar{D}_h$ onto $\mathcal{R}$ (actually just functions, since $\bar{D}_h$ is a finite set) with the norm

$$\|v\| = \sup\{|v(x)| : x\in\bar{D}_h\}. \tag{2.42}$$

Hence there exists a unique solution to (2.40).

Since we can express for any u_h^0 given the following:

$$u_h = \lim_{i\to\infty} u_h^i, \quad u_h^{i+1} = T_h(u_h^i), \quad i=0,1,\cdots,$$

we easily deduce (2.38), where $\|\cdot\|$ denotes the supremum norm in the corresponding space, i.e., for $\|f\|$ we take the supremum over $\bar{D}_h\times\Lambda$ or $\mathcal{R}_h^d\times\Lambda$.

To check (2.39), we use the discrete maximum principle on the function $w = \pm u_h - \|f\|\bar{u}_h$. □

Consider the extended finite difference operator $\tilde{L}_h(\lambda)$ given by

$$\begin{aligned}
&\tilde{L}_h(\lambda)(x,y) = h^{-1}\sum_{k=1}^{n}[\tilde{L}_h^+(k,\lambda)\varphi(x,y)+\tilde{L}_h^-(k,\lambda)\varphi(x,y)],\\
&h\tilde{L}_h^{\pm}(k,\cdot)\varphi(x,y) = p_k^{\pm}(x,y,h)[\varphi(x+\gamma_k^{\pm}(x,h), y+\gamma_k^{\pm}(y,h))-\varphi(x,y)]\\
&\qquad\qquad + q_k^{\pm}(x,y,h)[\varphi(x+\gamma_k^{\pm}(x,h), y)-\varphi(x,y)]\\
&\qquad\qquad + q_k^{\pm}(y,x,h)[\varphi(x,y+\gamma_k^{\pm}(y,h))-\varphi(x,y)],
\end{aligned} \tag{2.43}$$

where

$$
\begin{aligned}
&p_k^{\pm}(x, y, \lambda, h) = \beta_k^{\pm}(y, \lambda, h), \quad \wedge = \text{minimum}, \\
&q_k^{\pm}(x, y, \lambda, h) = \beta_k^{\pm}(x, \lambda, h) - \beta_k^{\pm}(x, \lambda, h) \wedge \beta_k^{\pm}(y, \lambda, h).
\end{aligned} \tag{2.44}
$$

Note that our choice implies that

$$
\tilde{L}_h(\lambda)[\varphi(x) + \psi(y)] = L_h(\lambda)\varphi(x) + L_h(\lambda)\psi(y), \tag{2.45}
$$

for any functions $\varphi(x)$, $\psi(y)$, and $L_h(\lambda)$ as in (2.11), (2.12). It is clear that

$$
\alpha_p(h) = \sup\{m(x-y, -p, h)\tilde{L}_h(\lambda)m(x-y, h): x, y \in \mathscr{R}_h^d, \lambda \in \Lambda\}, \tag{2.46}
$$

is finite, for h in $(0, 1]$, $0 < p \leqq 1$, and

$$
m(x, p, \varepsilon) = (\varepsilon + |x|^2)^{p/2}, \qquad x \in \mathscr{R}^d, \quad \varepsilon > 0. \tag{2.47}
$$

Suppose that (2.1), (2.16), and

$$
\begin{aligned}
&|\beta_k^{\pm}(x, \lambda, h) - \beta_k^{\pm}(y, \lambda, h)| \leqq K|x-y|, \\
&|\gamma_k^{\pm}(x, \lambda, h) - \gamma_k^{\pm}(y, \lambda, h)| \leqq Kh^{1/2}|x-y|, \\
&|q_k^{+}(x, y, h)\gamma_k^{+}(x, \lambda, h) + q_k^{-}(x, y, \lambda, h)\gamma_k^{-}(x, \lambda, h)| \leqq Kh|x-y|,
\end{aligned} \tag{2.48}
$$

for some constant $K > 0$, uniformly for the variables x, y in $\mathscr{R}_h^d$, λ in Λ, h in $(0, 1]$, $k = 1, 2, \cdots, n$, hold true. Then, for some constant C depending on the various constants of the hypotheses (2.1), (2.16), and (2.48), we have

$$
\alpha_p(h) \leqq \alpha_p + Ch^{p/2}, \qquad p > 0, \tag{2.49}
$$

where

$$
\begin{aligned}
\alpha_p = k \sup \Bigg\{ & |x-y|^{-2} \sum_{i=1}^{d} [g_i(x, \lambda) - g_i(y, \lambda)](x_i - y_i) \\
& + (p \vee 1)|x-y|^4 \sum_{k=1}^{n} \sum_{i,j=1}^{d} (x_i - y_i)(x_j - y_j)[\sigma_{ik}(x, \lambda) - \sigma_{jk}(y, \lambda)] \\
& \cdot [\sigma_{jk}(x, \lambda) - \sigma_{jk}(y, \lambda)] + |x-y|^{-2} \sum_{k=1}^{n} \sum_{i=1}^{d} [\sigma_{ik}(x, \lambda) - \sigma_{ik}(x, \lambda)]^2 : \\
& x \neq y' \text{ in } \mathscr{R}^d, \lambda \text{ in } \Lambda \Bigg\}.
\end{aligned} \tag{2.50}
$$

Note that $\alpha_p(h)$ and α_p vanish as p goes to zero. The condition (2.48) is almost equivalent to the Lipschitz condition of (2.1), i.e., that (2.16) and (2.48) imply (2.1) and (2.48) is expected to hold if we wish to insure (2.1).

THEOREM 2.2. *Under the assumptions* (2.4), (2.11), (2.12), (2.37), *and*

$$
\alpha > \max\{\alpha_p(h), \beta_p\} \quad \forall h \in (0, 1], \tag{2.51}
$$

the unique solution u_h *to the discrete* HJB *equation* (2.27) *satisfies*

$$
|u_h(x) - u_h(y)| \leqq K|x-y|^p \quad \forall x, y \in \bar{D}_h, \quad h \in (0, 1], \tag{2.52}
$$

for some constant K depending only on the various constants appearing in the hypotheses (2.4), (2.37), *and* (2.51).

Proof. Consider the function

$$
w(x, y) = u_h(x) - u_h(y) - Km(x-y, h, p),
$$

where $m(\cdot,\cdot,\cdot)$ is given by (2.47) and $K>0$ is a constant to be selected later. We want to show that $w(x,y)\leqq 0$ at any point x, y of $\bar{D}_h$, which implies (2.52), since $|x-y|\geqq h$, if $x\neq y$. Let (x_0, y_0) be a point in $\bar{D}_h\times\bar{D}_h$ where the function $w(x,y)$ attains its maximum value; such a point exists always. If either $x_0\in\Gamma_h$ or $y_0\in\Gamma_h$ then $w(x_0,y_0)\leqq 0$ provided $K\geqq\|f\|K(\bar{u}_h)$, the constant of (2.37). Herein we have used the estimate (2.39). Let us look at the case when x_0, y_0 belong to D_h.

By means of the discrete maximum principle for the extended operator $\tilde{L}_h(\lambda)$ we have $\tilde{L}_h(\lambda)w(x_0,y_0)\leqq 0$, which implies

$$L_h(\lambda)u_h(x_0)-L_h(\lambda)u_h(y_0)\leqq K\tilde{L}_h(\lambda)m(x_0-y_0,h,p)$$

after using (2.45). If we choose $0<\alpha_0\leqq\alpha-\alpha_p(h)$, for any h in $(0,1]$, then in view of (2.46) we get

$$\tilde{L}_h(\lambda)m(x_0-y_0,p,h)\leqq(\alpha-\alpha_0)m(x_0-y_0,p,h)$$

for any l in Λ. Since u_h satisfies the discrete HJB equation (2.27) at x_0 and y_0, we deduce

$$\begin{aligned}\alpha[u_h(x_0)-u_h(y_0)]&\leqq\sup\{|f(x_0,\lambda)-f(y_0,\lambda)|:\lambda\in\Lambda\}\\&\quad+K\sup\{\tilde{L}_h(\lambda)m(x_0-y_0,p,h):\lambda\in\Lambda\}\\&\leqq[K(f)+(\alpha-\alpha_0)K]m(x_0-y_0,p,h)\end{aligned}$$

where $K(f)$ is the constant of hypothesis (2.4). Hence if we take $K=\alpha_0^{-1}K(f)$, then $w(x_0,y_0)\leqq 0$, i.e.,

$$u_h(x)-u_h(y)\leqq Km(x-y,p,h)\quad\forall x,y\in\bar{D}_h,\quad h\in(0,1].$$

Thus, the estimate (2.52) follows. □

Remark 2.1. Note that in the assumption (2.51) we suppose implicitly that (2.16) and (2.48) hold true.

2.3. Main estimate. Let us look at the continuous time HJB equation (2.7), i.e.,

$$\alpha u=\inf\{L(\lambda)u+f(\cdot,\lambda):\lambda\in\Lambda\}\quad\text{in } D,\quad u=0\quad\text{on }\partial D, \tag{2.53}$$

where the differential operator is

$$L(\lambda)=\sum_{i,j=1}^{d}a_{ij}(\cdot,\lambda)\partial_{ij}+\sum_{i=1}^{d}a_i(\cdot,\lambda)\partial_i, \tag{2.54}$$

and we have identified the coefficients

$$\begin{aligned}&\sum_{k=1}^{n}\sigma_{ik}(x,\lambda)\sigma_{jk}(x,\lambda)=2a_{ij}(x,\lambda)\quad\forall x\in\bar{D},\quad\lambda\in\Lambda,\\&g_i(x,\lambda)=a_i(x,\lambda)\quad\forall x\in\bar{D},\quad\lambda\in\Lambda.\end{aligned} \tag{2.55}$$

Suppose that

$$\begin{aligned}&D\text{ is a bounded domain in }\mathcal{R}^d\text{ with smooth boundary }\partial D,\text{ say }C^{2,p}\text{ for}\\&\text{some }0<p\leqq 1,\end{aligned} \tag{2.56}$$

and

$$\begin{aligned}&\alpha\geqq 0,\text{ and for some }\nu_0>0\text{ we have }\nu_0|x|^2\leqq\textstyle\sum_{i,j=1}^{d}a_{ij}(x,\lambda)\xi_i\xi_j\leqq\nu_0^{-1}|\xi|^2,\\&\forall x\in\bar{D},\ \lambda\in\bar{D},\ \lambda\in\Lambda,\ \xi\in\mathcal{R}^d.\end{aligned} \tag{2.57}$$

It has been proven independently by Evans [7] and Krylov [14] (cf. Gilbarg and Trudinger [11]) that under the assumptions (2.56), (2.57), and

$$a_{ij},\ a_i,\ f\text{ are smooth, say }C^3\text{ in }x\text{ uniformly in }\lambda, \tag{2.58}$$

the HJB equation (2.53) has a unique classic solution, which is continuous on the closure $\bar{D}$ and its first- and second-order derivatives are Hölder-continuous for some exponent $0<p_0<1$ on the open domain D.

This result has been improved by Krylov [15] to show that under the same assumptions, the first- and second-order derivatives of the solution u are indeed Hölder-continuous on the closure $\bar{D}$.

Then an almost optimal result due to Safonov [34] provides an equivalent of Schauder estimates for HJB equations. Precisely, under the assumptions (2.56), (2.57), and

$$\|\varphi\|_{(p)} \leqq K \quad \forall \varphi = a_{ij}(\cdot, \lambda), a_i(\cdot, \lambda) \quad \forall \lambda \in \Lambda, \tag{2.59}$$

where $\|\cdot\|_{(p)}$ denotes the Hölder norm in $C^p(\bar{D})$, $0<p\leqq 1$, there exists a constant $p_0(\nu_0, d)$ in $(0, 1]$ such that the estimate

$$\|u\|_{(2+p)} \leqq C \sup\{\|f(\cdot, \lambda)\|_{(p)}: \lambda \in \Lambda\}, \tag{2.60}$$

holds for some constant C depending only in K, ν_0, p and the domain D, provided $0<p<p_0(\nu_0, d)$. Note that $\|\cdot\|_{(2+p)}$ denotes the Hölder norm in the space $C^{2,p}(\bar{D})$.

Another case in the quasilinear equation is

$$a_{ij}(x, \lambda) = a_{ij}(x) \quad \forall x \in \bar{D} \quad \forall \lambda \in \Lambda. \tag{2.61}$$

Thus we do not control the diffusion term. Under the conditions (2.56), (2.57), (2.59), and (2.61) the estimate (2.60) holds for every $0<p<1$ (cf. Ladyzhenskaya and Uraltseva [19]).

It is known (cf. Lions [20], Lions and Menaldi [21], Krylov [13]) that under the assumptions (2.1), (2.4), (2.9), and (2.10) the HJB equation (2.53) has a unique solution in some weak sense, e.g., either as the maximum subsolution with $L(\lambda)$ acting in the Schwarz distributions sense or as the unique viscosity solution. Moreover, if we denote by u_ε the solution of the HJB equation (2.53) with $L(\lambda)$ replaced by $L(\lambda)+\varepsilon\Delta$, Δ the Laplacian operator, then we can assert that

$$u_\varepsilon \in C^2(\bar{D}), \quad u_\varepsilon \to u \quad \text{in } C(\bar{D}), \tag{2.62}$$

where $C(\bar{D})$ is the space of continuous functions on $\bar{D}$.

For a smooth function φ, say $C^{2,p}(\bar{D})$, $0<p\leqq 1$, let us define $[\varphi]_p$ as the infimum of the set of all constant C satisfying

$$\begin{aligned} &\inf\{L(\lambda)\varphi(y): |y-x| \leqq C\sqrt{h}\} - Ch^{p/2} \leqq L_h(\lambda)\varphi(x) \\ &\qquad \leqq \sup\{L(\lambda)\varphi(y): |y-x| \leqq C\sqrt{h}\} + Ch^{p/2} \quad \forall h \in (0, 1], \end{aligned} \tag{2.63}$$

for any x in D_h, λ in Λ.

It is clear that $[\varphi]_p$ can be bounded by the constant C_φ of the estimate (2.17). However, occasionally we can do better:

(i) $[\varphi]_1$ is dominated by the bounds of the second-order derivatives of φ and the constants $C(g, \sigma)$, $K(g, \sigma)$, C of hypotheses (2.1), (2.16), provided $n=1$. This means that only one-dimensional Brownian motion is allowed, e.g., the system associated with an equation of order d perturbed by a white noise.

(ii) If $\sigma = \sigma(\lambda)$ and $n=1$, i.e., constant in x and only one Brownian motion, and

$$\sigma_{i1}\sigma_{j1}h = (\beta_1^+ + \beta_1^-)^{-1}\beta_1^+\beta_1^-(\gamma_{i1}^+ - \gamma_{i1}^-)(\gamma_{j1}^+ - \gamma_{j1}^-) \quad \forall i, j, \lambda,$$

then $[\varphi]_p$ is dominated by the p-Hölder norm of the first-order derivatives of φ and the constants $C(g, \sigma)$, $K(g, \sigma)$,

(iii) If $g=0$ and (ii) holds, then $[\varphi]_1$ is dominated by the bound of σ and does not depend on φ.

THEOREM 2.3. *Let the assumptions* (2.1), (2.4), (2.9)-(2.12), (2.16), (2.37), (2.48), *and*

$$\alpha > \max\{\alpha_p, \beta_p\}, \qquad 0 < p \leqq 1 \tag{2.64}$$

hold true. Then there exists a constant $C>0$ *depending only on the various constants of the hypotheses* (2.4), (2.37), (2.48), (2.64), *and* $[u]_p$ *as in* (2.63), *such that*

$$|u_h(x)-u(x)| \leqq Ch^{p/2} \quad \forall x \in \bar{D}_h, \quad h \in (0,1], \tag{2.65}$$

where u_h *is the solution of the discrete* HJB *equation* (2.27) *and* u *is the viscosity solution of the* HJB *equation* (2.53).

Proof. We remark that we are using (2.62) to suppose $[u]_p$ finite and defined as the limit of $[u_\varepsilon]_p$.

First, we will give a proof where the constant C of (2.65) depends on the C_φ, $\varphi = u$ of the convergence property (2.17), i.e., the p-Hölder norm of the second-order derivatives of u. This argument uses implicitly the discrete maximum principle in a way similar to Lions and Mercier [22].

Indeed, let us define the nonlinear resolvent operators

$$R(\varphi) = v \text{ iff } v=0 \text{ on } \partial D \text{ and } \alpha v = \varphi + \inf\{L(\lambda)v + f(\cdot,\lambda): \lambda \in \Lambda\} \text{ in } D, \tag{2.66}$$

and

$$\begin{aligned} &R_h(\varphi_h) = v_h \text{ iff } v_h = 0 \text{ on } \Gamma_h \text{ and} \\ &\alpha v_h = \varphi_h + \inf\{L_h(\lambda)v_h + f(\cdot,\lambda): \lambda \in \Lambda\} \text{ in } D_h. \end{aligned} \tag{2.67}$$

It is clear that if u_h and u denote the solutions to the HJB equations (2.27) and (2.53), then

$$u_h - u = R_h(0) - R(0) = R_h[R^{-1}(R(0))] - R_h[R_h^{-1}(R(0))],$$

where R^{-1} and R_h^{-1} are the inverse operators. By means of Theorem 2.1, the inequality (2.38) gives

$$\|R_h(\varphi) - R_h(\psi)\| \leqq \alpha^{-1}\|\varphi - \psi\| \quad \forall h \in (0,1],$$

for any functions φ, ψ and with $\|\cdot\|$ denoting the supremum norm on $\bar{D}_h$. Hence

$$\|u_h - u\| \leqq \alpha^{-1}\|R^{-1}(u) - R_h^{-1}(u)\|.$$

Since we can bound

$$\begin{aligned} |R^{-1}(u) - R_h^{-1}(u)| &= |\inf\{L(\lambda)u + f(\cdot,\lambda): \lambda \in \Lambda\} - \inf\{L_h(\lambda)u + f(\cdot,\lambda): \lambda \in \Lambda\}| \\ &\leqq \sup\{|L(\lambda)u - L_h(\lambda)u|: \lambda \in \Lambda\} \leqq C_u h^{p/2}, \end{aligned}$$

where C_u is the constant in (2.17), we deduce the estimate (2.65) with $C = \alpha^{-1}C_u$.

Next, to fully prove (2.65) we will show first that

$$|u(x)| \leqq C[\operatorname{dist}(x, \partial D)]^p \quad \forall x \in \bar{D}, \tag{2.68}$$

$$|u(x) - u(y)| \leqq K|x-y|^p \quad \forall x, y \in \bar{D}, \tag{2.69}$$

for some constants C, K depending only on the various constants of (2.4), (2.9), (2.10), and (2.64). To that effect, we construct a p-Hölder-continuous subsolution $\bar{u}$, i.e., a function $\bar{u}$ satisfying in a weak sense,

$$\begin{aligned} &L(\lambda)\bar{u} \leqq -1 + \beta_p \bar{u} \quad \text{in } D, \quad \bar{u} = 0 \quad \text{on } \partial D, \\ &|\bar{u}(x) - \bar{u}(y)| \leqq K|x-y|^p \quad \forall x, y \in \bar{D}, \end{aligned} \tag{2.70}$$

for some constants $K = K(\bar{u})$, $\beta_p \geqq 0$, as in (2.37). The maximum principle applied to the function

$$w(x) = \pm u(x) - \|f\| \bar{u}(x)$$

yields $w(x) \leqq 0$, i.e., (2.68).

To obtain (2.69) we proceed as in Theorem 2.2. We consider the extended differential operator

$$\begin{aligned} \tilde{L}(\lambda) = \frac{1}{2} \sum_{i,j=1}^{d} \sum_{k=1}^{n} [& \sigma_{ik}(x,\lambda)\sigma_{jk}(x,\lambda)\partial_{ij}^{x} + \sigma_{ik}(x,\lambda)\sigma_{jk}(y,\lambda)\partial_{ij}^{xy} \\ & + \sigma_{ik}(y,\lambda)\sigma_{jk}(x,\lambda)\partial_{ij}^{yx} + \sigma_{ik}(y,\lambda)\sigma_{jk}(y,\lambda)\partial_{ij}^{y} \\ & + \sum_{i=1}^{d} [g_i(x,\lambda)\partial_i^x + g_i(y,\lambda)\partial_i^y]], \end{aligned} \tag{2.71}$$

where ∂_{ij}^{x} and ∂_{ij}^{xy} denote the derivatives with respect to x_i, x_j and x_i, y_j, respectively. A simple calculation shows that

$$\tilde{L}(\lambda) m(x-y, p, \varepsilon) \leqq \alpha_p m(x-y, p, \varepsilon) \quad \forall x, y \in \mathcal{R}^d, \quad \lambda \in \Lambda, \tag{2.72}$$

any $\varepsilon > 0$ and with α_p being the constant (2.50). The function $m(\cdot, \cdot, \cdot)$ is given by (2.47).

Now, define the function

$$w(x, y) = u(x) - u(y) - Km(x-y, p, \varepsilon),$$

for some constant $K > 0$ to be selected. Let (x_0, y_0) be a point in $\bar{D} \times \bar{D}$ where $w(x, y)$ attains its maximum values. We want to prove that $w(x_0, y_0) \leqq 0$, which implies (2.69) as ε vanishes. In fact, if either $x_0 \in \partial D$ or $y_0 \in \partial D$, then $w(x_0, y_0) \leqq 0$ provided $K \geqq C(u)$, the constant in (2.68). So, if x_0, y_0 belong to D, then the maximum principle yields $\tilde{L}(\lambda) w(x_0, y_0) \leqq 0$, i.e.,

$$L(\lambda)u(x_0) - L(\lambda)u(y_0) \leqq K\tilde{L}(\lambda) m(x_0 - y_0, p, \varepsilon).$$

In view of the HJB equation (2.53) and the inequality (2.72) we get

$$\begin{aligned} \alpha[u(x_0) - u(y_0)] &\leqq \sup\{|f(x_0, \lambda) - f(y_0, \lambda)| : \lambda \in \Lambda\} \\ &\quad + K \sup\{\tilde{L}(\lambda) m(x_0 - y_0, p, \varepsilon) : \lambda \in \Lambda\} \\ &\leqq [K(f) + \alpha_p K] m(x_0 - y_0, p, \varepsilon), \end{aligned}$$

where $K(f)$ is the constant of hypotheses (2.4). Hence, take $K = (\alpha - \alpha_p)^{-1} K(f)$ to obtain $w(x_0, y_0) \leqq 0$.

Let us prove the estimate (2.65). To that effect, we consider the function

$$w(x, y) = u_h(x) - u(y) - C_1 m(x-y, p, h) - C_2 h^{p/2},$$

for any x, y in $\bar{D}_h$, and some constants $C_1, C_2 > 0$ to be selected. We want to show that at the point (x_0, y_0) in $\bar{D}_h \times \bar{D}_h$ where $w(x, y)$ attains its maximum value, we have $w(x_0, y_0) \leqq 0$, from which (2.65) follows immediately. Indeed if either $x_0 \in \Gamma_h$ or $y_0 \in \Gamma_h$ then $w(x_0, y_0) \leqq 0$, provided $C_1 = \max\{K(u_h), K(u)\}$, the p-Hölder constants given by (2.52) and (2.69). Actually, the constants in (2.39) and (2.68) suffice, i.e., p-Hölder constants near the boundary. When $x_0, y_0 \in D_h$, we can use the discrete maximum principle for the extended operator $\tilde{L}_h(\lambda)$, defined by (2.43), to get $\tilde{L}_h(\lambda) w(x_0, y_0) \leqq 0$, i.e.,

$$L_h(\lambda)u_h(x_0) - L_h(\lambda)u(y_0) \leqq C_1 \tilde{L}_h(\lambda) m(x_0 - y_0, p, h).$$

In view of the discrete HJB equation (2.27) we get

$$L_h(\lambda)u_h(x_0) - L_h(\lambda)u(y_0) \geqq \alpha u_h(x_0) - [L(\lambda)u(y) + f(y, \lambda)] + [L(\lambda)u(y) - L_h(\lambda)u(y_0)] + [f(y, \lambda) - f(x_0, \lambda)].$$

Thus, by taking the supremum over λ in Λ and y such that

$$|y - y_0| \leqq [u]_p \sqrt{h},$$

we deduce

$$\alpha[u_h(x_0) - u(y_0)] - [u]_p h^{p/2} - [\alpha K(u) + K(f)](1 + [u]_p^p) m(x_0 - y_0, p, h) \leqq \alpha_p(h) C_1 m(x_0 - y_0, p, h),$$

after using the definition (2.46) and (2.63), and the p-Hölder constants $K(f)$, $K(u)$ of (2.4), (2.69). Since we need only to show (2.65) for $h > 0$ sufficiently small, the hypothesis (2.64) and the inequality (2.49) permit us to choose

$$C_1 \geqq [\alpha - \alpha_p(h)]^{-1}(1 + [u]_p^p)[\alpha K(u) + K(f)],$$

$$C_2 = \alpha^{-1}[u]_p,$$

in order to have $w(x_0, y_0) \leqq 0$, i.e.,

$$u_h(x) - u(y) \leqq C m(x - y, p, h) \quad \forall x, y \in \mathcal{R}_h^d, \quad h \in (0, 1],$$

where $C = a + C_2$; this implies one side of (2.65).

By symmetry we obtain (2.65) after using the estimate (2.52) of Theorem 2.2. □

2.4. Final comments. Sometimes we need to discretize the set Λ, where the control takes values. In this case, a new term of the form

$$\sup\{\inf\{|l(x, \lambda) - l(x, \lambda')|: \lambda' \in \Lambda(h)\}: x \in \mathcal{R}^d, \lambda \in \Lambda\}, \tag{2.73}$$

for $l = f$, g_i, σ_{ik}, $i = 1, \cdots, d$, $k = 1, \cdots, n$, will appear in the right-hand side of the estimate (2.65). Here $\Lambda(h)$ is a discretization of Λ. Also, the constant α could be a function $\alpha(x, \lambda)$, for which the preceding results extend. If the domain D is unbounded and the data f has polynomial growth, then the solutions $u(x)$, $u_h(x)$ will have also polynomial growth, and some weight function is needed to obtain an estimate similar to (2.65) (cf. [24]).

We may be interested in the performance of the optimal control of the discrete problem, when suitably extended and applied to the initial problem. That issue is not considered here. However, the optimizer will face the problem of actually computing $u_h(x)$. In general, only an approximation $\tilde{u}_h(x)$ is computed and from that, a control policy $\tilde{\lambda}_h(\cdot)$ is derived. This $\tilde{\lambda}_h(\cdot)$ allows us to simulate a trajectory $\tilde{y}_h(\cdot)$. To this policy $(\tilde{\lambda}_h(\cdot), \tilde{y}_h(\cdot))$ a new cost $\hat{u}_h(x)$ is associated. Then, starting from (2.65) we need really to estimate $|\tilde{u}_h(x) - \hat{u}_h(x)|$. Again, this issue is not addressed here.

As mentioned in the theorems, the assumption on uniform ellipticity (2.57) is not required, at least explicitly. For instance, the case of a one-dimensional Brownian motion can be considered. This includes the control of a one-dimensional ordinary equation of order n, perturbed by a white noise.

We have assumed (2.10) for simplicity and to have the Dirichlet condition on the whole boundary ∂D. However, we need only to correctly identify the part Γ of the boundary where the diffusion process exists, and then we can use the technique described in this paper. This requires supposing that the operator $L(\lambda)$ is degenerate with constant order of degeneration, i.e., (2.21).

From the practical point of view, the estimate $h^{1/2}$ is not relevant, since better results are usually expected. However, this gives a precise relation between the grid for the space variable and the control variable, when $\Lambda(h)$ is used. The constant $\alpha_p(h)$ defined by (2.46) plays an important role in the stability of the numerical schemes. This is not obtained in classical schemes.

Perhaps the most interesting part is the fact that the finite difference operator (0.6) does not require any condition of "stability." It is stable in nature, and most estimates valid for the differential operator (0.5) have an equivalent in the discrete case.

In Bancora-Imbert, Chow, and Menaldi [1], the numerical solution of an optimal correction problem for a damped random linear oscillator is studied. The HJB equation takes the form of a variational inequality, namely,

$$\begin{aligned} &\partial_t u + Lu \geqq 0 \quad \text{in } \mathcal{R}_2 \times [0, T), \qquad -c \leqq \partial_2 u \leqq c \quad \text{in } \mathcal{R}_2 \times [0, T), \\ &(\partial_t u + Lu)(\partial_2 u + c)(\partial_2 u - c) = 0 \quad \text{in } \mathcal{R}_2 \times [0, T), \\ &u(\cdot, T) = f \quad \text{in } \mathcal{R}_2, \end{aligned} \tag{2.74}$$

where the differential operator is given by

$$Lu(x_1, x_2, t) = \tfrac{1}{2} r^2 \partial_2^2 u(x_1, x_2, t) - (px_2 + q^2 x_1)\partial_2 u(x_1, x_2, t) + x_2 u(x_1, x_2, t),$$

and r, p, q, c are constants; $r, q, c > 0$; and f is a given function. A precise algorithm is described and used there. The solution of the discrete problem is found as the common limit of two sequences, one decreasing and the other increasing. This allows us to bound the error and to give an almost optimal policy. We refer also to Sun and Menaldi [37], [25]. Note that in the case of (2.74), the solution u is Lipschitz-continuous together with its second derivative in the x_2 variable.

In a subsequent paper, the (quasi-) variational inequalities will be studied. It is well known that for those problems the solution is not smooth, i.e., the second derivative must have a jump. For that reason, only the second approach of Theorem 2.3, i.e., using $[u]_p$, seems to be appropriate. Perhaps a combination with finite elements of the type used by Menaldi and Rofman [26], [23] could be of some help.

Acknowledgments. The author thanks Professor P. L. Chow for the useful discussions on this work and the reviewers for the opportunity to improve the first version of the paper.

REFERENCES

[1] M. C. Bancora-Imbert, P. L. Chow, and J. L. Menaldi, *On the numerical approximations of an optimal correction problem*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 970-991.

[2] A. Bensoussan and J. L. Lions, *Applications des inequations variationnelles en controle stochastique*, Dunod, Paris, 1978. English translation, North-Holland, Amsterdam, 1982.

[3] D. P. Bertsekas and S. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[4] Capuzzo-Dolcetta, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., 10 (1983), pp. 367-377.

[5] I. Capuzzo-Dolcetta and H. Ishii, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161-181.

[6] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comput., 43 (1984), pp. 1-19.

[7] L. C. Evans, *Classical solutions of Hamilton-Jacobi-Bellman equation for uniformly elliptic operators*, Trans. Amer. Math. Soc., 275 (1983), pp. 245-255.

[8] M. Falcone, *A numerical approach to the infinite horizon problem*, Appl. Math. Optim., 15 (1987), pp. 1-14.

[9] W. FLEMING AND R. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] I. I. GIHKMAN AND A. V. SKOROKHOD, *Controlled Stochastic Processes*, Springer-Verlag, New York, 1979.

[11] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Second edition, Springer-Verlag, New York, 1983.

[12] R. GONZALEZ AND E. ROFMAN, *On deterministic control problems: an approximate procedure for the optimal cost, Parts* I *and* II, SIAM J. Control Optim., 23 (1985), pp. 242-285.

[13] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1981.

[14] ———, *Boundedly nonhomogeneous elliptic and parabolic equations in a domain*, Math. USSR Izv., 22 (1984), pp. 67-97.

[15] ———, *On estimates for the derivatives of solutions of nonlinear parabolic equations*, Soviet Math. Dokl., 29 (1984), pp. 14-17.

[16] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[17] ———, *Approximation and Weak Convergence Methods for Random Processes with Application to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[18] H. J. KUSHNER AND A. KLEINMAN, *Accelerated procedures for the solution of discrete Markov control problems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 147-152.

[19] O. A. LADYZHENSKAYA AND N. N. URALTSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[20] P. L. LIONS, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Part* 2: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229-1276.

[21] P. L. LIONS AND J. L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations, Parts* I *and* II, SIAM J. Control Optim., 20 (1982), pp. 58-95.

[22] P. L. LIONS AND B. MERCIER, *Approximation numerique des equations de Hamilton-Jacobi-Bellman*, RAIRO Anal. Numer., 14 (1980), pp. 369-393.

[23] J. L. MENALDI, *Sur l'approximation numerique des inequations variationnelles elliptiques*, Publicazioni di Laboratorio di Analisi Numerica di C.N.R. 140, University of Pavia, Italy, 1976.

[24] ———, *Probabilistic View of Estimates for Finite Difference Methods*, IMA Preprint Services, 266, 1986. See also Math. Notae, to appear.

[25] ———, *Bounded Variation Control of a Damped Linear Oscillator Under Random Disturbances, in Stochastic Differential Systems, Stochastic Control and Applications*, P. L. Lions and W. H. Fleming, eds., IMA 10, Springer-Verlag, Berlin, New York, 1988, pp. 373-394.

[26] J. L. MENALDI AND E. ROFMAN, *Sur les problems variationnels non coercifs et l'equation du transport*, Lecture Notes in Mathematics 606, Springer-Verlag, Berlin, New York, 1976, pp. 202-209.

[27] ———, *An algorithm to compute the viscosity solution Hamilton-Jacobi-Bellman equation*, in Theory and Applications of Nonlinear Control Systems, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 461-467.

[28] E. PARDOUX AND D. TALAY, *Discretization and simulation of stochastic differential equations*, Acta Appl. Math., 3 (1985), pp. 23-47.

[29] M. L. PUTERMAN, *On the convergence of policy iteration of controlled diffusions*, Working Paper 590, Faculty of Commerce, University of British Columbia, 1978.

[30] M. L. PUTERMAN AND S. L. BRUMELLE, *On the convergence of policy iteration in stationary dynamic programming*, Math. Oper. Res., 4 (1979), pp. 60-69.

[31] J. P. QUADRAT, *Existence de solution et algorithme de resolutions numeriques de problems stochastiques degenerees ou non*, SIAM J. Control Optim., 18 (1980), pp. 199-226.

[32] S. ROSS, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.

[33] THEOSYS, *Commande Optimale de Systemes Stochastiques*, RAIRO Automatique, 18 (1984), pp. 225-250.

[34] M. V. SAFONOV, *On the classical solution of Bellman's elliptic equation*, Soviet Math. Dokl., 30 (1984), pp. 482-485.

[35] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Probab. Appl., 1 (1956), pp. 261-290.

[36] P. E. SOUGANIDIS, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1-43.

[37] M. SUN AND J. L. MENALDI, *Monotone control of a damped oscillator under random perturbations*, IMA J. Math. Control Inform., 5 (1988), pp. 169-186.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 608-630, May 1989

009

ASYMPTOTIC PROPERTIES OF OPTIMAL SOLUTIONS IN PLANAR DISCOUNTED CONTROL PROBLEMS*

FRITZ COLONIUS†‡ AND MALTE SIEVEKING†

Abstract. The classical Poincaré-Bendixson Theorem on limit sets of solutions of planar differential equations is generalized to solutions of planar optimal control problems maximizing a discounted present value that does not depend explicitly on the control function.

Key words. infinite horizon optimal control, Poincaré-Bendixson Theorem, optimal resource management

AMS(MOS) subject classifications. 49D50, 49A99, 34C25, 92A15

1. Introduction. The main result of this paper is a generalization of the classical Poincaré-Bendixson Theorem for the following class of optimal control problems (P) (with $n=2$):

$$\text{(1.1)} \qquad \text{Maximize} \quad \int_0^\infty e^{-\delta t} R(x(t))\, dt,$$

subject to

$$\text{(1.2)} \qquad \dot{x}_j(t) = x_j(t)\left[f_0^j(x(t)) + \sum_{i=1}^m u_i(t) f_i^j(x(t)) \right] \quad \text{a.a. } t \in \mathbb{R}_+, \quad j = 1, \cdots, n,$$

$$\text{(1.3)} \qquad u(t) = (u_i(t)) \in \Omega \subset \mathbb{R}^m \quad \text{a.a. } t \in \mathbb{R}_+,$$

$$\text{(1.4)} \qquad x(0) = x \in \mathbb{R}_+^n;$$

here $\delta > 0$ and $R : \mathbb{R}_+^n \to \mathbb{R}$, $f_i = (f_i^j) : \mathbb{R}_+^n \to \mathbb{R}$ are locally Lipschitz continuous, the set Ω of control values is compact and convex, and the control functions are chosen in

$$\text{(1.5)} \qquad U_{\text{ad}} = U_{\text{ad}}(\mathbb{R}_+) := \{u : \mathbb{R}_+ \to \Omega, \text{ measurable}\}.$$

Thus we consider discounted optimal control problems where the integral of the performance index does not depend explicitly on the control, and the system equation has the "ecological form" (1.2) with control appearing linearly. Our original motivation for considering asymptotic properties of optimal solutions comes from bioeconomics. Here the study of such problems is often decomposed into two parts.

First an optimal equilibrium point e is searched for and then a determination optimal approach path from the initial point x_0 to e is tried (compare, e.g., Clark [5, p. 317]). This approach is justified in the case of a single-state variable ($n=1$), since here, in general, bounded solution $x(\cdot)$ of (P) converge monotonically to an optimal equilibrium as t tends to $+\infty$ (see Theorem 2.7 below). For two state variables ($n=2$) the classical Poincaré-Bendixson Theorem describes the asymptotic behavior of the special class of (uncontrolled) differentiable dynamical systems; here the limit set $\omega(x)$ of a trajectory either is a periodic trajectory or consists of trajectories connecting

* Received by the editors January 21, 1987; accepted for publication (in revised form) August 17, 1988. This research was supported by a grant from Stiftung Volkswagenwerk.

† Fachbereich Mathematik, Johann Wolfgang Goethe Universität, Frankfurt, Federal Republic of Germany.

‡ University of Bremen, D-28 Bremen 33, Federal Republic of Germany.

equilibria. The Poincaré-Bendixson theorem generalizes within the framework of local (nondifferentiable) dynamical systems (see Hajek [10]). However, optimal solutions of (P) are not, in general, unique (cf. the examples given in § 6 below). Hence they need not form a local dynamical system. Nevertheless, the present paper shows that the Poincaré-Bendixson Theorem can be generalized to optimal control problems of the type above. The problem of nonuniqueness must be met at the following crucial steps in the reconstruction of the classical argument for proving the Poincaré-Bendixson Theorem.

(1) We want to form Jordan arcs from parts of solutions; however, solutions may be self-intersecting. This problem is solved via the optimality principle: if an optimal solution x returns at time $t_2 > t_1$ into the same state as at time t_1, then it is also optimal to run through the same piece of trajectory $x|[t_1, t_2]$ again and again. The solution obtained this way is optimal and periodic after time t_1. This is our justification for studying only the limit sets of nonself-intersecting solutions.

(2) To construct "flow boxes" at nonequilibria we cannot rely on a local parallelization theorem such as in dynamical systems. To prove existence of transversal sections and appropriately defined "flow boxes" we use that the integrand in (1.1) does not depend explicitly on the control u. Sometimes more general problems can be reduced to this form (cf. Remarks 2.12, 4.8). The possibility of defining optimality by means of such a functional has been exploited by Clark in a number of resource management problems.

The literature on the asymptotic behavior of optimal solutions for (P) concentrates mainly on establishing sufficient conditions for convergence to equilibrium. We only mention Arrow [2], Rockafellar [15], [16], Feinstein and Luenberger [8], and Feinstein and Oren [9]. The convexity assumptions made here are quite restrictive and are usually not satisfied in resource management. Haurie [12], [13] relaxes the convexity condition, such that they are, e.g., applicable to Volterra-Lotka equations. However, he must assume existence of optimal equilibria (with additional properties).

Oscillatory behavior of optimal solutions is often attributed to nonlinear cost effects and to age structure (Clark [5, pp. 166, 293], Deklerk and Gatto [7]). In § 5 we present an example that possesses neither of these attributes.

For a problem arising in economics, Benhabib and Nishimura [4] analyze the optimality system resulting from the Pontryagin maximum principle. Taking the discount rate δ as a bifurcation parameter, they show that Hopf bifurcations occur. The corresponding periodic solutions are optimal due to convexity assumptions.

The paper is organized as follows. Section 2 contains the basic assumptions and what is needed later about convergent subsequences of solutions and their limit sets. Furthermore the key lemma about transversal segments is proved as well as the existence of "flow boxes." Section 3 is a study of optimal equilibria. As a consequence of Pontryagin's maximum principle it is shown that in "general" there are only finitely many optimal equilibria, and a sufficient condition for attractivity of optimal equilibria is established. The main result is Theorem 3.5, which settles a case in the Poincaré-Bendixson Theorem. Section 4 contains the proof of the Poincaré-Bendixson Theorem for nonself-intersecting solutions. Section 5 discusses resource management problems. A predator-prey system where the predator is subject to harvesting is analyzed. As a consequence of the Poincaré-Bendixson Theorem, there are optimal solutions having as limit set an optimal periodic solution. Section 6 discusses nonuniqueness arising when an optimal periodic trajectory does not contain an optimal equilibrium in its interior as well as nonuniqueness in an example of a symmetric system of two harvested competing species. Here nonuniqueness stems from the bifurcation of behavior in the nonharvested system.

2. Limit sets and flow boxes. We will denote solutions of (1.2), (1.4) by $\varphi(t, x, u)$, $t \geqq 0$, and always assume global existence of $\varphi(\cdot, x, u)$ on $\mathbb{R}_+$ (uniqueness follows from local Lipschitz continuity). Let $\varphi(x, u) := \{\varphi(t, x, u): t \in \mathbb{R}_+\}$ and denote the value of (1.1) corresponding to (x, u) by $V(x, u)$. If not specified otherwise, convergence in U_{ad} means weak convergence in the L_2-sense on compact intervals. We will also use

$$U_{\text{ad}}(\mathbb{R}) := \{u : \mathbb{R} \to \Omega, \text{measurable}\}.$$

Throughout this paper, we assume that the following hypothesis is satisfied:

(2.1) For every compact subset $K \subset \mathbb{R}^n_+$, the set $\{\varphi(t, x, u): t \in \mathbb{R}_+, x \in K, u \in U_{\text{ad}}\}$ is bounded.

DEFINITION 2.1. A pair $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ is called optimal if for all $v \in U_{\text{ad}}$ we have

$$V(x, u) \geqq V(x, v).$$

A pair $(e, u^e) \in \mathbb{R}^n_+ \times \Omega$ is called an optimal equilibrium, if $e = \varphi(t, e, u^e)$ for all $t \in \mathbb{R}_+$ and the pair (e, u^e) is optimal (here u^e is identified with the constant control in U_{ad} with value u^e). For an optimal pair (x, u), we let $V(x) := V(x, u)$.

Remark 2.2. The notion of optimality above keeps the initial point $x(0) = x$ fixed and considers only the effect of different control actions v.

Remark 2.3. Frequently it will—instead of (2.1)—be sufficient that for a fixed (optimal) pair (x, u), we have that $\varphi(x, u) \subset \mathbb{R}^n_+$ is bounded.

Remark 2.4. Let $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ be given and suppose that for some $t > 0$, we have $\varphi_j(t, x, u) = 0$ for all $j \in J \subset \{1, 2, \cdots, n\}$. Define

$$\psi_i(s, x, u) = \begin{cases} 0, & i \in J, \\ \varphi_i(s, x, u), & i \notin J, \end{cases}$$

for s in a neighborhood of t. Then ψ also solves (1.2) and $\psi(t, x, u) = \varphi(t, x, u)$. Hence by the uniqueness of solutions of ordinary differential equations $\varphi = \psi$ in a neighborhood of t. Hence either $\varphi_j(s, x, u) = 0$ for all $s \geqq 0$ or $\varphi_j(s, x, u) > 0$ for all $s \geqq 0$. Therefore none of the species can become extinct in finite time and for any $J \subset \{1, \cdots, n\}$ $\{y \mid y_j = 0, j \in J\} \cap \mathbb{R}^n_+ = \mathbb{R}^{(J)}_+$ is invariant and the restriction of the system to $\mathbb{R}^{(J)}_+$ is a system of the same form.

LEMMA 2.5. *Suppose* $x^k \to x^0$ *in* $\mathbb{R}^n_+$ *and* $u^k \to u^0$ *in* U_{ad}. *Then* $\varphi(\cdot, x^k, u^k) \to \varphi(\cdot, x^0, u^0)$ *uniformly on bounded intervals and* $V(x^k, u^k) \to V(x^0, u^0)$.

Proof. The first assertion follows in a standard way from Gronwall's inequality. For the second one, take $\varepsilon > 0$. Then for T and k large enough and $x^k(t) := \varphi(t, x^k, u^k)$, $t \in \mathbb{R}_+$, $k = 0, 1, 2, \cdots$,

$$\begin{aligned} \left| V(x^0, u^0) - \int_0^\infty e^{-\delta t} R(x^k(t))\, dt \right| &\leqq \left| V(x^0, u^0) - \int_0^T e^{-\delta t} R(x^0(t))\, dt \right| \\ &\quad + \left| \int_0^T e^{-\delta t} [R(x^0(t)) - R(x^k(t))]\, dt \right| \\ &\quad + \left| \int_T^\infty e^{-\delta t} R(x^k(t))\, dt \right| \\ &\leqq 3\varepsilon, \end{aligned}$$

using the first assertion and (2.1).

COROLLARY 2.6. *Let* $(x^k, u^k) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ $(k \in \mathbb{N})$ *be optimal and* $(x^k)_k$ *bounded. Then there are a subsequence* (x^{k_i}, u^{k_i}) $(i \in \mathbb{N})$ *and an optimal* (x, u) *such that* $\lim_i \varphi(\cdot, x^{k_i}, u^{k_i}) = \varphi(\cdot, x, u)$ *locally uniformly and* $\lim_i u^{k_i} = u$ *in* U_{ad}.

Proof. Existence of a subsequence (x^{k_i}, u^{k_i}) converging to (x, u) follows from boundedness. Now let $v \in U_{\text{ad}}$. Then, by optimality of (x^{k_i}, u^{k_i}),

$$V(x, u) = \lim V(x^{k_i}, u^{k_i}) \geqq \lim V(x^{k_i}, v) = V(x, v).$$

This proves optimality of (x, u).

Let $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ be optimal and define $x(\cdot) := \varphi(\cdot, x, u)$.

THEOREM 2.7. (1) *Suppose that $n = 1$ and $x(t)$ is neither increasing nor decreasing; then $\alpha < \beta$ exist such that each $e \in (\alpha, \beta)$ is an optimal equilibrium.*

(2) *If $e = \lim_{t\to\infty} x(t)$ then e is an optimal equilibrium.*

Proof. (1) If $x(t)$ is neither increasing nor decreasing, there exist $r < s < t$ such that $x(r) = x(t)$, $x(s) \neq x(r)$. We may choose s such that either $x(s) = \min x([r, t])$ or $x(s) = \max x([r, t])$, say $x = \max x([r, t])$. Choose any $b \in (\alpha, \beta) = (x(r), x(s))$. There is a first instant $r_\varepsilon > r$ such that $x(r_\varepsilon) = b$, a first instant $s_\varepsilon > r_\varepsilon$ such that $x(s_\varepsilon) = b + \varepsilon$, and a first instant $t_\varepsilon > s_\varepsilon$ such that $x(t_\varepsilon) = b$. Let s'_ε be the last instant $< t_\varepsilon$ such that $x(s'_\varepsilon) = b + \varepsilon$. Then define

$$u_\varepsilon(\sigma) = u(r_\varepsilon + \sigma) \qquad \text{for } 0 \leqq \sigma \leqq s_\varepsilon - r_\varepsilon,$$

$$u_\varepsilon(s_\varepsilon - r_\varepsilon + \sigma) = u(s'_\varepsilon + \sigma) \quad \text{for } 0 \leqq \sigma \leqq t_\varepsilon - s_\varepsilon.$$

This way $u_\varepsilon(\sigma)$ is defined for $0 \leqq \sigma \leqq s_\varepsilon - r_\varepsilon + t_\varepsilon - s'_\varepsilon = \pi_\varepsilon$. Now extend u_ε to obtain a periodic function on $\mathbb{R}_+$ with period π_ε. Define x_ε in the same way as u_ε using $x(t)$ instead of $u(t)$. Then x_ε satisfies $\dot{x}_\varepsilon(t) = f(x_\varepsilon(t), u_\varepsilon(t))$ almost everywhere on $\mathbb{R}_+$ and $(x_\varepsilon, u_\varepsilon)$ is a solution of (1.1)-(1.4) with $x_\varepsilon(0) = b$. Note that for all $t \geqq 0$ we have $|x_\varepsilon(t) - b| \leqq \varepsilon$. Let $\varepsilon_n > 0$ tend to zero. Then Corollary 2.6 implies that b is an optimal equilibrium.

(2) Suppose $e = \lim_{t\to\infty} x(t)$. For $n \in \mathbb{N}$ put $x_n(t) = x(t+n)$, $u_n(t) = u(t+n)$. Then (x_n, u_n) solves (1.1)-(1.4) with $x_n(0) = x(n)$. Hence e is an optimal equilibrium by Corollary 2.6.

Next we introduce the central notions of this paper.

DEFINITION 2.8. For $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ define the omega limit set $\omega(x, u)$ by

$$\omega(x, u) := \{y \in \mathbb{R}^n : \text{there exist } t_k \in \mathbb{R}_+ \text{ such that } t_k \to \infty \text{ and } \varphi(t_k, x, u) \to y\} = \bigcap_{n\in\mathbb{N}} \operatorname{cl}\{\varphi(t, x, u) : t \geqq n\}. \tag{2.2}$$

For $I = \mathbb{R}_+$ or $I = \mathbb{R}$, we call $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}(I)$ an optimal I-solution if the corresponding solution $\varphi(\cdot, x, u)$ of (1.2) exists on I and for all $t \in I$

$$V(\varphi(t, x, u), u(t + \cdot)) = V(\varphi(t, x, u)).$$

Frequently, we call optimal $\mathbb{R}_+$-solutions simply optimal. If (x, u) is an optimal $\mathbb{R}$-solution, define the alpha limit set $\alpha(x, u)$ by

$$\alpha(x, u) := \bigcap_{n\in\mathbb{N}} \operatorname{cl}\{\varphi(t, x, u) : t \leqq -n\}. \tag{2.3}$$

Finally define for optimal (x, u)

$$\hat{\omega}(x, u) := \{(y, v) : (y, v) \text{ is an optimal } \mathbb{R}\text{-solution and there are } t_k \in \mathbb{R}_+ \text{ such that } t_k \to \infty \text{ and } \varphi(t_k + \cdot, x, u) \to \varphi(\cdot, y, v) \text{ locally uniformly on } \mathbb{R} \text{ and } u(t_k + \cdot) \to v \text{ in } U_{\text{ad}}\}. \tag{2.4}$$

DEFINITION 2.9. A subset L of $\mathbb{R}^n_+$ is called (*positively*) *invariant* if for all $y \in L$ there is an optimal ($\mathbb{R}_+$-) $\mathbb{R}$-solution (y, v) with $\varphi(\mathbb{R}, y, v) \subset L$ (respectively, $\varphi(\mathbb{R}_+, y, v) \subset L$).

PROPOSITION 2.10. *Let (x, u) be optimal. Then $\omega(x, u)$ is nonvoid, compact, and connected. For every $y \in \omega(x, u)$ there is $v \in U_{ad}$ such that $(y, v) \in \hat{\omega}(x, u)$ and $\varphi(\mathbb{R}, y, v) \subset \omega(x, u)$. In particular, $\omega(x, u)$ is invariant.*

Proof. Using (2.1), we see that $\omega(x, u)$ is nonvoid, connected, and compact. Let $y \in \omega(x, u)$. Then there are $t_k \to \infty$ with $\varphi(t_k, x, u) \to y$. By Corollary 2.6, we can, without loss of generality, assume that $s \to u(t_k + s)$, $s \in \mathbb{R}_+$, converges in U_{ad} to some $u_0 \in U_{ad}$ and $s \to \varphi(t_k + s, x, u)$, $s \in \mathbb{R}_+$, converges uniformly on bounded intervals to $\varphi(\cdot, y, u_0)$. Taking again, if necessary, subsequences, $s \to u(t_k - 1 + s)$, $s \in \mathbb{R}_+$, converges weakly on bounded intervals to $u_{-1} : [-1, \infty) \to \Omega$ and $s \to \varphi(t_k - 1 + s, x, u)$ converges uniformly on bounded intervals to $\varphi(\cdot, y, u_{-1}) : [-1, \infty) \to \mathbb{R}^n$ with $u_0 = u_{-1}$ and $\varphi(\cdot, y, u_0) = \varphi(\cdot, y, u_{-1})$ on $[0, \infty)$. By successively taking subsequences of (t_k) we obtain sequences $u_{-n} : [-n, \infty) \to \Omega$, $\varphi(\cdot, y, u_{-n}) : [-n, \infty) \to \mathbb{R}^n$ with $u_{-n} = u_{-n+1}$ on $[-n+1, \infty)$ and

$$\frac{d}{dt}\varphi(t, y, u_{-n}) = f(\varphi(t, y, u_{-n}), u_{-n}(t)),$$

$$V(\varphi(-n, y, u_{-n}), u_{-n}(-n + \cdot)) = V(\varphi(-n, y, u_{-n})).$$

Defining

$$v(t) = u_{-n}(t) \quad \text{on } [-n, \infty),$$

we obtain an optimal $\mathbb{R}$-solution (y, v).

The following lemma is our key for the construction of local transversal sections.

LEMMA 2.11. *Let L be a compact positively invariant set and $R(e) = \sup\{R(x) \mid x \in L\}$ for some $e \in L$. Then one of the following conditions is satisfied:*

(i) *e is an optimal equilibrium;*

(ii) *L contains a point x^0 with $0 \notin f(x^0, \Omega)$.*

Proof. If (ii) is violated there is $v^e \in \Omega$ such that $f(e, v^e) = 0$. By invariance of L we find $v \in U_{ad}$ such that (e, v) is optimal with $\varphi(\mathbb{R}_+, e, v) \subset L$. Hence

$$V(e) = \int_0^\infty e^{-\delta t} R(\varphi(t, e, v))\, dt \leqq \int_0^\infty e^{-\delta t} R(e)\, dt$$

$$= \int_0^\infty e^{-\delta t} R(\varphi(t, e, v^e))\, dt = V(e, v^e).$$

Thus (e, v^e) is an optimal equilibrium, i.e., (i) holds.

Remark 2.12. In § 5, we will consider a two-dimensional problem from resource management ($n = 2$), where the integrand of the performance criterion depends also on u. However, the problem can be transformed into one in which in the *interior* of $\mathbb{R}^2_+$ we obtain a criterion of the form (1.1) ($R(x)$ becomes unbounded for $x \to \partial\mathbb{R}^2_+$). In fact, Lemma 2.11 remains true here, since it holds in the following general situation.

Suppose (1.1) is replaced by

$$\int_0^\infty e^{-\delta t}\left[g_0(x(t)) + \sum_{i=1}^m u_i(t) g_i(x(t))\right] dt, \tag{2.5}$$

with $g_i : \mathbb{R}^2_+ \to \mathbb{R}$, $i = 0, 1, \cdots, m$, locally Lipschitz continuous, and the following condition holds:

(2.6) There is a continuous function $R : \operatorname{int} \mathbb{R}^2_+ \to \mathbb{R}$ such that

(α) $\int_0^\infty e^{-\delta t} R(\varphi(t, a, u))\, dt$ converges for every $a \in \operatorname{int} \mathbb{R}^2_+$, $u \in U_{ad}$

to a real number $V_R(a, u)$.

(β) A pair $(a, u) \in \operatorname{int} \mathbb{R}^2_+ \times U_{\text{ad}}$ is optimal if and only if $V_R(a, u) \geqq V_R(a, v)$ for all $v \in U_{\text{ad}}$.

First observe that Corollary 2.6, Proposition 2.10, and Theorem 2.7 remain true for the criterion (2.5). If $L \subset \operatorname{int} \mathbb{R}^2_+$ this follows from Lemma 2.11. Otherwise, $L \cap \partial \mathbb{R}^2_+ \neq \emptyset$; suppose, for example, $L_1 = L \cap \pi_1^{-1}(0) \neq \emptyset$ where $\pi(x_1, x_2) = x_1$. By Remark 2.4, the restriction of the system to $\pi_1^{-1}(0)$ is well defined and L_1 is compact and invariant for this system. Since the restricted system again is of the form (2.5), (1.2)-(1.4), Theorem 2.7 yields the assertion.

We first consider case (ii) of Lemma 2.11, and show that it translates into a geometric condition.

DEFINITION 2.13. Let $x^0 \in \mathbb{R}^n$, $l: \mathbb{R}^n \to \mathbb{R}$ linear and $\alpha > 0$. If

$$lf(y, u) > \alpha$$

for all y in a neighborhood W of x^0 and all $u \in \Omega$, then

$$S := W \cap l^{-1}(x^0)$$

is called a local transversal section through x^0.

PROPOSITION 2.14. *Suppose that $0 \notin f(x^0, \Omega)$. Then x^0 possesses a local transversal section. Hence a compact positively invariant set L either contains an optimal equilibrium or a point possessing a local transversal segment.*

Proof. In view of Lemma 2.11, we only have to show the first assertion. If $0 \notin f(x^0, \Omega)$, then by the Hahn-Banach Theorem this assertion follows, since $f(x^0, \Omega)$ is compact and convex.

Obviously, trajectories "can cross a local transversal section only from one side." The next result presents an important consequence from the existence of a local transversal section.

We need the following definition.

DEFINITION 2.15. Let S be a local transversal section through x^0, and let $V_1 \subset V_0$ be neighborhoods of x^0. Then the triple (V_0, V_1, S) is called a flow box around x^0, if it has the following property:

If $\varphi(\cdot, x^0, u)$ satisfies

$$\varphi(t_0, x^0, u) \notin V_0, \quad \varphi(t_1, x^0, u) \in V_1, \quad \varphi(t_2, x^0, u) \notin V_0$$

for some $0 \leqq t_0 < t_1 < t_2$, then there exists $t \in (t_0, t_2)$ such that $\varphi(t, x^0, u) \in S$ and $\varphi(s, x^0, u) \in V_0$ for all s between t and t_1.

THEOREM 2.16. *Let S be a local transversal section through x^0. Then there are neighborhoods V_0 and V_1 of x^0 such that (V_0, V_1, S) is a flow box around x^0.*

Proof. There exist a linear map $l: \mathbb{R}^n \to \mathbb{R}$, a constant $\alpha > 0$, and a neighborhood W of x^0 such that $S \supset W \cap l^{-1}(x^0)$ and

$$l(f(y, v)) > \alpha \quad \text{for all } y \in W, \quad v \in \Omega.$$

Choose a ball $V_0 = B(r_0, x^0)$ around x^0 with radius $r_0 > 0$ such that $V_0 \subset W$ and put $c := \sup \{|f(y, u)| \mid y \in V_0, v \in \Omega\}$. Then choose $r_1 \in (0, r_0)$ so small that

$$lz - \alpha/2c(r_0 - r_1) \leqq ly \leqq lz + \alpha/2c(r_0 - r_1) \tag{2.7}$$

for all z, $y \in V_1 = B(r_1, x^0)$. We have for $t > t' \geqq 0$:

$$\varphi(t, x, u) = \varphi(t', x, u) + \int_{t'}^{t} f(\varphi(s, x, u), u(s))\, ds$$

and hence

$$l(\varphi(t, x, u)) = l(\varphi(t', x, u)) + \int_{t'}^{t} lf(\varphi(s, x, u), u(s))\, ds$$
$$\geqq l(\varphi(t', x, u)) + \alpha(t - t')$$

provided that $\varphi(s, x, u) \in W$, $t' \leqq s \leqq t$. Without loss of generality, we may assume

$$\varphi(s, x, u) \in V_0 \quad \text{for all } t_0 \leqq s \leqq t_2$$

replacing, if necessary, t_0 by the last time before t_1 at which $\varphi(t, x, u)$ is in the complement of V_0 and t_2 by the first time after t_1 at which $\varphi(t, x, u)$ leaves V_0. We have

$$r_0 - r_1 \leqq |\varphi(t_1, x, u) - \varphi(t_0, x, u)| \leqq c(t_1 - t_0),$$
$$r_0 - r_1 \leqq |\varphi(t_2, x, u) - \varphi(t_1, x, u)| \leqq c(t_2 - t_1).$$

If $l\varphi(t_0, x, u) \leqq lx^0 \leqq l\varphi(t_2, x, u)$, or $l\varphi(t_2, x, u) \leqq lx^0 \leqq l\varphi(t_0, x, u)$, the assertion follows by continuity of $t \to lf(t, x, u)$. Hence we only have to consider the following two cases.

Case 1. $\dot{l}x^0 < \min\{l\varphi(t_0, x, u), l\varphi(t_2, x, u)\}$. Here $l\varphi(t_1, x, u) \geqq l\varphi(t_0, x, u) + \alpha(t_1 - t_0) > lx^0 + \alpha/c(r_0 - r_1)$, contradicting (2.7) for $y = \varphi(t_1, x, u)$.

Case 2. $lx^0 > \max\{l\varphi(t_0, x, u), l\varphi(t_2, x, u)\}$. Here $l\varphi(t_2, x, u) \geqq l\varphi(t_1, x, u) + a(t_2 - t_1) > l\varphi(t_1, x, u) + \alpha/c(r_0 - r_1)$, again contradicting (2.7).

3. Optimal equilibria. In this section we first characterize optimal equilibria by necessary optimality conditions. It turns out that "in general" only finitely many optimal equilibria exist. Strong additional assumptions ensure that optimal equilibria in a limit set are already reached in finite time. Furthermore, limit sets $\omega(x, u)$ reduce to a single optimal equilibrium provided that $\omega(x, u)$ consists of equilibria only and contains at most finitely many optimal equilibria.

First we discuss the following problem:

$$\text{Maximize (2.5)} \quad \text{subject to (1.2)-(1.4)}$$

where Ω is a rectangle in $\mathbb{R}^2$ (in fact, the "ecological form" of (1.2) is not needed in this section, if not stated otherwise).

Abbreviate

$$g(x, u) = g_0(x) + \sum_{i=1}^{m} u_i g_i(x), \qquad f(x, u) = f_0(x) + \sum_{i=1}^{m} u_i g_i(x).$$

For any equilibrium $e = (x_1, x_2)$, there are the two equations (for x_1, x_2, u_1, u_2) defining an equilibrium, namely

$$o = f(x, u). \tag{3.1}$$

To derive a second set of equations we shall use Pontryagin's maximum principle (cf. Halkin [11]). Write

$$H = \lambda_0 e^{-\delta t} g + \lambda \cdot f,$$
$$\dot{\lambda}(t) = -\lambda_0 e^{-\delta t} g_x - f_x^t \lambda \quad \text{(adjoint equation)}.$$

Here

$$f = \begin{pmatrix} f^1 \\ f^2 \end{pmatrix} \quad \text{and} \quad f_x^t = \begin{pmatrix} f^1_{x_1} & f^2_{x_1} \\ f^1_{x_2} & f^2_{x_2} \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}.$$

Thus,

$$H = \lambda_0 e^{-\delta t} g_0 + \lambda_1 f_0^1 + \lambda_2 f_0^2 + u_1(\lambda_0 e^{-\delta t} g_1 + \lambda_1 f_1^1 + \lambda_2 f_1^2) + u_2(\lambda_0 e^{-\delta t} g_2 + \lambda_1 f_2^1 + \lambda_2 f_2^2).$$

Put $\mu = e^{\delta t}\lambda$. Then $\dot{\mu} = \delta\mu + e^{\delta t}\dot{\lambda}$. Hence the adjoint equation reads

$$\dot{\mu} = -\lambda_0 g_x + (\delta I - f_x^t)\mu,$$

and

$$H = h(t, x, \lambda) + u_1(\lambda_0 g_1 + \mu_1 f_1^1 + \mu_2 f_1^2) + u_2(\lambda_0 g_2 + \mu_1 f_2^1 + \mu_2 f_2^2).$$

Pontryagin's maximum principle implies that $(\lambda_0, \lambda(t) \neq 0$ for all $t \geqq 0$ and H attains its maximum over Ω in (u_1, u_2). We may assume that $\lambda_0 = 0$ or $\lambda_0 = 1$.

Now we discuss the possible numbers of optimal equilibria. There are three cases:

Case 1. $u = (u_1, u_2)$ is one of the corners of Ω.

Case 2. u lies in the relative interior of one of the edges of Ω.

Case 3. $u \in \operatorname{int} \Omega$.

Case 1. Recall that there are only four corners of Ω.

Case 2. One equation for u is given by the condition that u lies on one of the edges of Ω. Furthermore the derivative of H in direction, say $v = (v_1, v_2)$ (parallel to the edge of Ω containing u), vanishes, i.e.,

$$(\lambda_0 g_1 + \mu_1 f_1^1 + u_2 f_1^2)v_1 + (\lambda_0 g_2 + \mu_1 f_2^1 + \mu_2 f_2^2)v_2 = 0 \quad \text{for all } t.$$

Thus with $\varphi := (v_1 f_1^1 + v_2 f_2^1,\ v_1 f_1^2 + v_2 f_2^2)$

$$\varphi\mu = -\lambda_0(g_1 v_1 + g_2 v_2) \quad \text{for all } t. \tag{3.2}$$

Insertion into the adjoint equation yields

$$0 = \varphi\dot{\mu} = \varphi(-\lambda_0 g_x + (\delta I - f_x^t)\mu) \quad \text{or} \quad \varphi(\delta I - f_x^t)\mu = \lambda_0 \gamma_x \quad \text{for all } t. \tag{3.3}$$

If φ and $\varphi(\delta I - f_x^t)$ are linearly dependent we obtain with (3.1), the assumption that u lies on an edge of Ω and

$$\det(\varphi', (\delta I - f_x^t)\varphi') = 0, \tag{3.4}$$

four equations for the unknowns x_1, x_2, u_1, u_2. If φ and $\varphi(\delta I + f_x^t)$ are linearly independent, (3.2) and (3.3) imply that μ is constant and that $\lambda_0 = 1$. We assume $\det(\delta I - f_x^t) \neq 0$. Then by the adjoint equation and (3.2)

$$\varphi(\delta I - f_x^t)^{-1} g_x = -g_1 v_1 + -g_2 v_2, \tag{3.5}$$

and again we obtain four equations for x_1, x_2, u_1, u_2.

Case 3. Put

$$F = \begin{pmatrix} f_1^1 & f_1^2 \\ f_2^1 & f_2^2 \end{pmatrix}.$$

Then

$$-\lambda_0 \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = F\mu.$$

Suppose $\det F \neq 0$. Then

$$\mu = -\lambda_0 F^{-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}$$

and it follows that $\lambda_0 = 1$, and $\mu = 0$. The adjoint equation yields

$$0 = g_x + (\delta I - f_x^t)\mu \quad \text{and} \quad g_x = (\delta I - f_x^t) F^{-1} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix}. \tag{3.6}$$

Hence together with (3.1) we obtain four equations for (x_1, x_2, u_1, u_2).

Now suppose $\det F = 0$. By introducing new control variables we can eliminate one control variable in the system equation and proceed with the discussion as in Case 1 or Case 2 above.

We formulate the conclusion of this discussion in the following remark.

Remark 3.1. Consider problem (2.5), (1.2)-(1.4) with $n=m=2$ and $\Omega=[0, U_1]\times[0, U_2]$. Then every optimal pair (x, u) such that $u\in\Omega$, $0=f(x, u)$ must satisfy four equations in the unknowns x_1, x_2, u_1, u_2. In concrete examples, these equations may serve to compute all candidates for optimal equilibria (recall that the maximum principle is only a set of necessary conditions). On the other hand, these equations justify the statement that "in general" there exist at most finitely many optimal equilibria. We shall use this as a hypothesis further below in this section and in § 4.

We proceed to analyze finite-time reachability properties of optimal equilibria in limit sets.

DEFINITION 3.2. An equilibrium e is called strongly optimal if the constant function $x(t)=e$ is the unique optimal trajectory for start in e.

LEMMA 3.3. *Let (x, u) be optimal and suppose that e is a strongly optimal equilibrium in $\omega(x, u)$. Then for every $T>0$ and every neighborhood V of e there exists a neighborhood U of e such that $\varphi(t, x, u)\in U$ implies $\varphi([t, t+T], x, u)\subset V$.*

Proof. Assume, contrary to the assertion, that there exist a neighborhood V of e, $T>0$, and $t_n\to\infty$ with $\varphi(t_n, x, n)\to e$ and $\varphi(t_n+s_n, x, u)\notin V$ for some $s_n\in[0, T]$. Without loss of generality, $s_n\to s\in[0, T]$. We may assume that $\varphi(\cdot, \varphi(t_n, x, u), u(t_n+\cdot))$ converges uniformly on bounded intervals to an optimal $\varphi(\cdot, e, v)$. Since $\varphi(s, e, v)\notin V$ and $\varphi(0, e, v)=e$, this contradicts strong optimality of e.

THEOREM 3.4. *Let $(x, u)\in\mathbb{R}^n_+\times U_{\mathrm{ad}}$ be optimal for* (1.1)-(1.4). *Let $e\in\omega(x, u)$ with the following*:

(i) *e is a strongly optimal equilibrium in* $\operatorname{int}\mathbb{R}^n_+$;

(ii) *There is $u^e\in\operatorname{int}\Omega$ with $f(e, u^e)=0$;*

(iii) *$m\geqq n$ and $f_1(e), \cdots, f_n(e)$ are linearly independent*;

(iv) *R is a C^2-function in a neighborhood of e and $R'(e)=0$, $R''(e)$ is negative definite.*

Then for all $t>0$ sufficiently large $\varphi(t, x, u)=e$.

Proof. (a) Suppose $\psi: V\to\mathbb{R}^n$ is a coordinate change defined on an open neighborhood V of e. If $x(t)$ is in V and satisfies $\dot{x}(t)=f_0(x(t))+\sum_{i=1}^n u_i(t)f_i(x(t))$, then $y(t)=\psi(x(t))$ satisfies $\dot{y}(t)=\dot{\psi}(x(t))(f_0\circ\psi^{-1})y(t)+\sum_{i=1}^n u_i(t)\dot{\psi}(x(t))(f_i\circ\psi^{-1})(y(t))$, which again is a system of equations of the type we are considering. Obviously our assumptions (i)-(iv) carry over. By (iv) and according to the Morse Lemma there is a coordinate change ψ such that $R\circ\psi^{-1}(x)=-\sum_{i=1}^n x_i^2$ for all x in a neighborhood W of $\psi(e)$.

Hence we may without loss of generality assume $R(e+x)=R(e)+\sum_{i=1}^n(x_i-e_i)^2$ in a ball $V(e, r)$ of center e and radius r.

(b) By (ii), (iii) and the implicit function theorem r may be chosen so small that a smooth function $u: V(0, r)\times V(e, r)\to\Omega$ exists such that for all $(y, x)\in V(0, r)\times V(e, r)$ we have $y=f(x, u(x, y))$. In fact, if $F(x)$ is the matrix with columns $f_1(x), \cdots, f_n(x)$ then

$$u(x, y)=F^{-1}(x)(y-f_0(x)).$$

In particular, if $x(t)=a+t(e-a)$, $\dot{x}(t)=e-a=f(x(t), u(x(t), e-a))$ provided $|e-a|<r$. Hence $x(t)=\varphi(t, a, u)$ is an admissible solution of our system at least up to $t=1$ $(u(t)=u(x(t), e-a))$.

(c) We now assume $e\notin\varphi(x, u)$ and try to reach a contradiction to (i). See Fig. 3.1. Choose $T>3$. According to Lemma 3.3 there is by (i) $t>0$ such that $\varphi((t, t+T), x, u)\subset V(e, r)$. By our assumption in (a) there is a first time $s_1\in(0, 1)$ such that for all $s\in(s_1, 1]$

$$R(x_0(s))>R(\varphi(t+s, x, u))$$

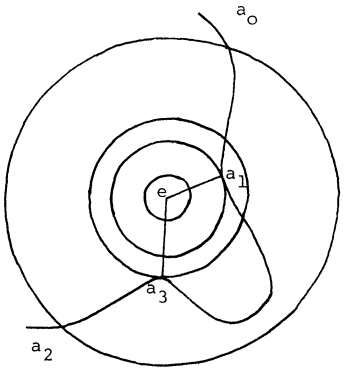

FIG. 3.1. $\hat{\varphi}$ *visits* a_0, a_1, e, a_3, a_2.

where

$$x_0(s) = a_0 + s(e - a_0), \quad a_0 = \varphi(t, x, u).$$

Put

$$a_1 = \varphi(t + s_1, x, u),$$

$$x_1(s) = a_1 + s\frac{|e - a_0|}{|e - a_1|}(e - a_1), \qquad 0 \leqq s \leqq \frac{|e - a_1|}{|e - a_0|} = s_2;$$

then

$$R(x_1(t)) > R(\varphi(t + s_1 + s, x, u)), \qquad 0 \leqq s \leqq s_2.$$

Put

$$a_2 = \varphi(t + T, x, u),$$

$$x_2(s) = e + s(a_2 - e), \qquad 0 \leqq s \leqq 1 = s_3.$$

Note that $s_1 + s_2 + s_3 < 3 < T$.

There is a last time $s_4 \in (0, s_3)$ such that

$$R(x_2(s)) > R(\varphi(t + T - s_3 + s, x, u)), \qquad s \in (0, s_4).$$

Put

$$a_3 = \varphi(t + T - s_3 + s_4, x, u),$$

$$x_3(s) = e + s\frac{|a_2 - e|}{|a_3 - e|}(a_3 - e), \qquad 0 \leqq s \leqq \frac{|a_3 - e|}{|a_2 - e|} = s_4.$$

Then

$$R(x_3(s)) > R(\varphi(t + T - s_3 + s, x, u)), \qquad s \in (0, s_4).$$

Now we combine the x_j's to build a solution that performs better than $\varphi(\cdot, x, u)$. Put

$$\hat{\varphi}(s) = \begin{cases} \varphi(s, x, u), & 0 \leqq s \leqq t + s_1, \\ x_1(s - t - s_2), & t + s_1 \leqq s \leqq t + s_1 + s_2, \\ e, & t + s_1 + s_2 \leqq s \leqq t + T - s_3, \\ x_3(s - t - T + s_3), & t + T - s_3 \leqq s \leqq t + T - s_3 + s_4, \\ \varphi(s, x, u), & t + T - s_3 + s_4 \leqq s. \end{cases}$$

The definition makes sense since $T > s_1 + s_2 + s_3$. Using (b) we may find $v \in U_{\text{ad}}$ such that $\hat{\varphi}(s) = \varphi(s, x, v)$. But

$$\sigma_1 = t + s_1 < t + T - s_3 + s_4 = \sigma_2$$

and

$$R(\varphi(s, x, v)) > R(\varphi(s, x, u)) \quad \text{for } s \in (\sigma_1, \sigma_2),$$
$$R(\varphi(s, x, v)) = R(\varphi(s, x, u)) \quad \text{for } s \notin (\sigma_1, \sigma_2).$$

Therefore $V(x, v) > V(x, u)$ in contradiction to the optimality of (x, u).

Next we give a sufficient condition for optimal solutions to converge to a single optimal equilibrium. This result will be used substantially in the next section.

THEOREM 3.5. *Let* $(x, u) \in \mathbb{R}^n_+ \times U_{\text{ad}}$ *be optimal for* (1.1)-(1.4). *Suppose the following*:

(i) $\omega(x, u)$ *contains at most finitely many optimal equilibria.*

(ii) $\omega(x, u) \subset E = \{y \in \mathbb{R}^n_+ \mid$ *there exists* $v \in \Omega$ *with* $o = f(y, v)\}$.

Then $\omega(x, u)$ *consists of a single optimal equilibrium.*

Proof. Suppose that

$$\# \omega(x, u) \geqq 2.$$

There is $e \in \omega(x, u)$ and a sequence of points $e_n \in \omega(x, u)$ such that $\lim_n e_n = e$ and

$$a = \sup \{R(y) \mid y \in \omega(x, u)\} = \lim_n R(e_n).$$

There is $\tilde{e} \in \omega(x, u)$ such that

$$b = R(\tilde{e}) < a.$$

Otherwise all points in $\omega(x, u)$ would be optimal equilibria and by connectedness $\omega(x, u)$ would contain infinitely many optimal equilibria contrary to (i). We choose numbers b_0, b_1, b_2 with

$$a > b_0 > b_1 > b_2 > b$$

such that there is at most one optimal equilibrium e_1 in $\omega(x, u)$ with $R(e_1) \geqq b_2$.

Choose $(t_k) \subset \mathbb{R}_+$ with $t_k \to \infty$ and $\varphi(t_k, x, u) \to e$. For every $k \in \mathbb{N}$ there are $s_i^k \geqq 0$, $i = 0, 1, 2$, with

$$s_2^k := \inf \{s \geqq 0 \colon R(\varphi(t_k + s, x, u)) < b_2\},$$
$$s_1^k := \sup \{s \leqq s_2^k \colon R(\varphi(t_k + s, x, u)) > b_1\},$$
$$s_0^k := \sup \{s \leqq s_1^k \colon R(\varphi(t_k + s, x, u)) > b_0\}.$$

We may assume that the functions

$$\varphi(t_k + s_0^k + \cdot, x, u)$$

converge uniformly on every bounded interval to an optimal trajectory $\varphi(\cdot, y, v) \subset \omega(x, u)$. Then $y \in E$ and $R(y) = b_0$. Next we show the following: The sequence $t_2^k = s_2^k - s_0^k$, $k \in \mathbb{N}$, is bounded. Otherwise we might assume that $t_2^k < t_2^{k+1}$ and $t_2^k \to \infty$. The functions

$$R(\varphi(t_k + s_0^k + t, x, u)), \qquad t \in [0, t_2^k]$$

have values in

$$V(b_0, b_2) := \{z \colon b_0 \geqq R(z) \geqq b_2\}.$$

Hence also $\varphi(\cdot, y, v)$, $\omega(y, v) \subset V(b_0, b_2)$.

Since $\omega(y, u) \subset E$, Lemma 2.11 implies that $\omega(y, v)$ contains an optimal equilibrium. This contradicts the choice of b_2.

Thus (t_2^k) is bounded and, considering subsequences, we may even assume that $t_2^k \to t_2$ and $t_1^k = s_1^k - s_0^k \to t_1$ with $0 \leqq t_1, t_2 < \infty$. Hence,

$$R(y) = R(\varphi(0, y, v)) = b_0,$$
$$R(\varphi(t, y, v)) \leqq b_0 \quad \text{for } t \in [0, t_1],$$
$$R(\varphi(t, y, v)) \leqq b_1 \quad \text{for } t \in [t_1, t_2],$$
$$R(\varphi(t, y, v)) \leqq a \quad \text{for } t \in [t_2, \infty).$$

We obtain

$$\int_0^\infty e^{-\delta t} R(\varphi(t, y, v))\, dt = \int_0^{t_1} + \int_{t_1}^{t_2} + \int_{t_2}^\infty \leqq \frac{b_0}{\delta}(1 - e^{t_1 \delta}) + \frac{b_1}{\delta}(e^{-t_1\delta} - e^{-t_2\delta}) + \int_{t_2}^\infty.$$

Since $y \in E$ and (y, v) optimal

$$\frac{1}{\delta} R(y) = \frac{b_0}{\delta} \leqq \int_0^\infty e^{-\delta t} R(\varphi(t, y, v))\, dt;$$

hence

$$b_0 \leqq b_0 - b_0\, e^{-t_1\delta} + b_1\, e^{-t_1\delta} - b_1\, e^{-t_2\delta} + \delta \int_{t_2}^\infty,$$

or

$$b_0 \leqq b_1(1 - e^{(t_1 - t_2)\delta}) + \delta\, e^{t_1\delta} \int_{t_2}^\infty R(\varphi(t, y, v))\, dt.$$

Note that the right-hand side is constructed independently of b_0. Hence if $a = +\infty$ we may let b_0 tend to $a = +\infty$, thus obtaining a contradiction since $\int_{t_2}^\infty$ converges to a finite value. Hence we may assume that a is finite and thus

$$b_0 \leqq b_1(1 - e^{(t_1 - t_2)\delta} + a\, e^{(t_1 - t_2)\delta}.$$

Letting b_0 tend to a, we find

$$a(1 - e^{(t_1 - t_2)\delta}) \leqq b_1(1 - e^{(t_1 - t_2)\delta}),$$

leading to the contradiction $a \leqq b_1$ since $\varphi(t_2, y, v) = b_2$, $\varphi(t_1, y, v) = b_1$, and hence $t_1 < t_2$.

Using similar arguments as in the proof above, we can show the analogous result for α-limit sets.

THEOREM 3.6. *Let the assumptions of Theorem* 3.5 *be satisfied for* $\alpha(x, u)$ *instead of* $\omega(x, u)$. *Then* $\alpha(x, u)$ *consists of a single optimal equilibrium.*

Remark 3.7. Let $n = 2$, replace (1.1) by (2.5), and suppose that (2.6) holds. Let $(x, u) \in \mathbb{R}_+^2 \times U_{\text{ad}}$ be optimal. Then when we assume (i), (ii), a slight change in the proof of Theorem 3.5 shows that either $\omega(x, u)$ consists of a single optimal equilibrium or is contained in the boundary of $\mathbb{R}_+^2$. Furthermore the following holds: Let $(y, u) \in \hat{\omega}(x, u)$ with $\omega(y, v) \subset \partial\mathbb{R}_+^2$ such that $\omega(y, v)$ contains only a finite number of optimal equilibria. Then $\omega(y, v)$ consists of a single optimal equilibrium.

Proof. Suppose the assertion is false. Then $\# \omega(y, v) \geqq 2$, and since $\omega(y, v)$ is connected, $\omega(y, v)$ contains infinitely many points. By assumption there is $z = (z_1, z_2) \in \omega(y, v)$, which is not an optimal equilibrium, say with $z_2 = 0$. There is $(z, w) \in \hat{\omega}(x, u)$. Since the first component $\varphi(\cdot, z, w)$ is not constant, it is by Theorem 2.7, say, increasing

(if it is decreasing, analogous arguments will apply). By Remark 2.4, the second component $\varphi_2(t, z, w)$ vanishes for all t. Let $t > 0$ such that the segment between z and $z' = \varphi(t, z, w)$ does not contain an optimal equilibrium. Since $z, z' \in \omega(x, u)$ there are t_n, $s_n \geqq 0$ such that

$$\lim_n t_n = +\infty, \quad \lim_n \varphi(t_n, x, u) = z', \quad \lim_n \varphi(t_n + s_n, x, u) = z$$

and for all $n \in \mathbb{N}$, $t \in [0, s_n]$,

$$z_1 \leqq \varphi_1(t_n + t, x, u) \leqq z_1', \qquad 0 \leqq \varphi_2(t_n + t, x, u) \leqq 1/n.$$

We may assume that the sequence of functions $t \to \varphi(t_n + t, x, u)$ converges locally uniformly to a function $\varphi(\cdot, z', w')$ such that $(z', w') \in \hat{\omega}(x, u)$. If $\lim s_n = +\infty$, then $\omega(z', w') \subset [z_1, z_1'] \times \{0\}$. By Theorem 2.7, $\omega(z', w')$ contains an optimal equilibrium, in contradiction to our assumption.

Therefore a subsequence of (s_n) converges to some $s \in [0, \infty)$. Obviously, $\varphi(s, z', w') = z$. Define

$$w''(\sigma) = w(\sigma) \qquad \text{for } 0 \leqq \sigma \leqq t,$$

$$w''(\sigma) = w'(\sigma - t) \qquad \text{for } t < \sigma \leqq s + t,$$

and extend w'' periodic on $\mathbb{R}_+$ with period $s + t$. Then (z, w'') is optimal and periodic. Now consider (z, w'') as an optimal pair with respect to the restriction of our system to $\mathbb{R}_+ \times \{0\} \simeq \mathbb{R}_+^1$ (Remark 2.4). Since $\varphi(\cdot, z, w'')$ is neither increasing nor decreasing, Theorem 2.7 implies the existence of infinitely many optimal equilibria contrary to the assumption.

4. Poincaré–Bendixson Theorem. The analysis of this section is restricted to two-dimensional systems (i.e., $n = 2$). Our final result, Theorem 4.6, is a generalization of the classical Poincaré-Bendixson Theorem. If we drop the assumption of that theorem that $\varphi(\cdot, x, u)$ is nonself-intersecting, we obtain an optimal periodic solution in a trivial way according to the following proposition.

PROPOSITION 4.1. *Suppose that $(x, u) \in \mathbb{R}_+^n \times U_{\text{ad}}$ is optimal and $\varphi(\cdot, x, u)$ intersects itself, i.e., there are $T_2 > T_1 \geqq 0$ with $\varphi(T_1, x, u) = \varphi(T_2, x, u)$. Then there is $\hat{u} \in U_{\text{ad}}$ such that $(x, \hat{u})$ is optimal and $\varphi(T_1 + s, x, u) = \varphi(T_1 + k(T_2 - T_1) + s, x, \hat{u})$ for $s \in [0, T_2 - T_1]$, $k \in \mathbb{N}$.*

Proof. Define $\hat{u}(t) = u(t)$ for $t \in [0, T_1]$,

$$\hat{u}(T_1 + k(T_2 - T_1) + t) = u(T_1 + t) \quad \text{for } t \in [0, T_2 - T_1], \quad k \in \mathbb{N}.$$

Then the assertion follows since final segments of optimal solutions are optimal.

We call a solution $(x, \hat{u})$ with the property above finally periodic.

For the reader's convenience, we cite the following classical theorem (see, e.g., Beck [3, Cor. C.23]), which will be used frequently.

JORDAN'S CURVE THEOREM. *Let J be a Jordan curve in $\mathbb{R}^2$ (i.e., a homeomorphism from the circle into $\mathbb{R}^2$). Then $\mathbb{R}^2 \setminus \operatorname{Im} J$ has two components, one of which is bounded (called* ins J*) and the other one (called* outs J*) is unbounded. Each one has boundary* Im J *and is simply connected.*

Since the orientation does not concern us, we identify J with its image.

LEMMA 4.2. *Let $(x, u) \in \mathbb{R}_+^2 \times U_{\text{ad}}$ and suppose that the corresponding trajectory $\varphi(\cdot, x, u)$ is nonself-intersecting. Then a local transversal section S has at most one point in common with $\omega(x, u)$. For optimal $\mathbb{R}$-solutions it follows also that S has at most one point in common with $\alpha(x, u)$.*

Proof. As in the theory of uncontrolled differential equations (see, e.g., [1, Lemma 24.1] or [14]) we prove the following: Let (x_i) be a sequence of points in $\varphi(x, u) \cap S$. If (x_i) is increasing on $\varphi(\cdot, x, u)$, then it is also on S. Now suppose that $y_1, y_2 \in \omega(x, u) \cap S$ and $y_1 \neq y_2$. Let U_j be disjoint neighborhoods of y_j, $j = 1, 2$. Then there exists a sequence $t_k \to \infty$ such that

$$\varphi(t_{2k+1}, x, u) \in U_1 \quad \text{and} \quad \varphi(t_{2k}, x, u) \in U_2, \quad k \in \mathbb{N}.$$

By Theorem 2.16, we may choose $U^j = V_1^j$, where (V_0^j, V_1^j, S) are flow boxes around y_j, $j = 1, 2$, and $V_0^1 \cap V_0^2 = \emptyset$. Then there exists a sequence (s_k), $s_k \to \infty$, with

$$\varphi(s_{2k+1}, x, u) \in U_1 \cap S, \qquad \varphi(s_{2k+1}, x, u) \in U_2 \cap S.$$

This contradicts the assertion above.

The same arguments apply to a α-limit sets of optimal $\mathbb{R}$-solutions.

PROPOSITION 4.2. *Let $(x, u) \in \mathbb{R}_+^2 \times U_{\text{ad}}$ be optimal and suppose that $\varphi(\cdot, x, u)$ is nonself-intersecting. Let (y, v) be an optimal $\mathbb{R}$-solution with $\varphi(\mathbb{R}, y, v) \subset \omega(x, u)$. Then $\omega(y, v)$ and $\alpha(y, v)$ consist of equilibria only or $\varphi(\cdot, y, v)$ intersects itself in a point z possessing a local transversal section S.*

Proof. Suppose that $\omega(y, v)$ contains a point z which is not an equilibrium. Then z possesses a local transversal section S by Lemma 2.11. Using a flow box around z we find that $\varphi(y, v) \cap S \neq \emptyset$. Since $\varphi(y, v)$, $\omega(y, v) \subset \omega(x, u)$ and $z \in \omega(y, v) \cap S \subset \omega(x, u) \cap S$ this implies by Lemma 4.2 that $S \cap \omega(x, u) = \{z\}$, and hence $\{z\} = \varphi(y, v) \cap \omega(y, v)$. Thus there is $T_1 \geq 0$ such that $\varphi(T_1, y, v) = z$. Since z is not an equilibrium, there is a neighborhood V_0 of z and $s > T_1$ with $\varphi(s, y, v) \notin V_0$. Using a flow box (V_0, V_1, S) around z, we find a $T_2 > s$ with $\varphi(T_2, y, v) \in S$. Hence $\varphi(T_2, y, v) = \varphi(T_1, y, v) = z$. Thus $\varphi(\cdot, y, v)$ intersects itself in z.

We prepare the proof of the next proposition by the following lemma.

LEMMA 4.4. *Let (C_n) be a decreasing sequence of closed sets in $\mathbb{R}^q$. Define $C := \bigcap_n C_n$, let $n_k \to \infty$, and*

$$D := \{y \in \mathbb{R}^q : \text{there exist } x_{n_k} \text{ with } x_{n_k} \in \partial C_{n_k} \text{ and } x_{n_k} \to y\}.$$

Then $\partial C = D$.

Proof. Suppose $y \in D$, i.e., there are (x_{n_k}) with $x_{n_k} \in \partial C_{n_k}$ and $x_{n_k} \to y$. Let $B(y, \varepsilon)$ be the ball with center y and radius $\varepsilon > 0$. Then for k large enough $x_{n_k} \in B(y, \varepsilon)$. Since $x_{n_k} \in \partial C_{n_k}$ there is $y_{n_k} \in B(y, \varepsilon) \setminus C_{n_k} \subset B(y, \varepsilon) \setminus C$. Since $\varepsilon > 0$ is arbitrary, $y \notin \operatorname{int} C$. Since $C_{n_k} \subset C_{n_l}$ for $k > l$ it follows that $x_{n_k} \in C_{n_l}$ for $k > l$ and, since C_{n_l} is closed, $y \in C_{n_l}$ for all l. Hence $y \in C \setminus \operatorname{int} C = \partial C$.

Conversely suppose that $y \in \partial C$ and note $C = \bigcap_k C_{n_k}$. Then for every $\varepsilon > 0$ there exists $z \in B(y, \varepsilon) \setminus C$. Hence there is n_k such that $y \in B(y, \varepsilon) \setminus C_{n_k}$. Suppose that $B(y, \varepsilon) \cap \partial C_{n_k} = \emptyset$. Then

$$B(y, \varepsilon) = (B(y, \varepsilon) \setminus C_{n_k}) \cup (B(y, \varepsilon) \cap \operatorname{int} C_{n_k}).$$

Since $B(y, \varepsilon)$ is connected and $z \in B(y, \varepsilon) \setminus C_{n_k} \neq \emptyset$ we conclude that $\emptyset = B(y, \varepsilon) \cap \operatorname{int} C_{n_k} = B(y, \varepsilon) \cap C_{n_k} \ni y$. This contradiction shows that there exist $y_{n_k} \in B(y, \varepsilon) \cap \partial C_{n_k}$. Evidently, $\lim y_{n_k} = y$, and hence $y \in D$.

PROPOSITION 4.5. *Let $(x, u) \in \mathbb{R}_+^2 \times U_{\text{ad}}$ be optimal and assume that $\varphi(\cdot, x, u)$ is nonself-intersecting. Suppose that there are $(y, v) \in \hat{\omega}(x, u)$ and $T_2 > T_1$ with*

$$\varphi(T_1, y, v) = \varphi(T_2, y, v) =: z,$$

with z possessing a local transversal section S. Then

$$\omega(x, u) = \varphi([T_1, T_2], y, v).$$

Proof. Since $z \in \omega(x, u)$ we can use a flow box around z to construct inductively a sequence of numbers t_n such that t_{n+1} is the first instant t after t_n with $\varphi(t, x, u) \in S$. Then for all n we have $t_n < t_{n+1}$, $t_n \to \infty$, and $\varphi(t_n, x, n) \to z$. Now define the Jordan arc Γ_n to consist of $\varphi([t_n, t_{n+1}], x, u)$ and the segment on S between $\varphi(t_n, x, n)$ and $\varphi(t_{n+1}, x, u)$. There are two cases.

Case 1. For all n $\operatorname{ins} \Gamma_n \supset \operatorname{ins} \Gamma_{n+1}$.

Case 2. For all n $\operatorname{outs} \Gamma_n \supset \operatorname{outs} \Gamma_{n+1}$.

Let us first consider Case 1. Put $C := \bigcap_n \operatorname{cl} \operatorname{ins} \Gamma_n$. By Lemma 4.4, $\partial C = \omega(x, u)$. Now let $l \in \mathbb{N}$ be arbitrary and consider a flow box (V_0, V_1, S) around z such that V_0 is a ball around z with radius $1/l$. The set V_1 contains a ball around z of positive radius r. Since $(y, v) \in \hat{\omega}(x, u)$, there is $t > 0$ such that

$$|\varphi(t + T_1 + s, x, u) - \varphi(T_1 + s, y, v)| < r, \qquad 0 \leqq s \leqq T_2 - T_1.$$

By the flow box property we may follow $\varphi(t + T_1 + s, x, u)$ starting with $s = 0$ and without leaving V_0 until we reach $t + T_1 + s = t_{l_n}$. Applying the same argument to the instant $t + T_2 - T_1$ we find that the part of $\varphi([t_{n_l}, t_{n_{l+1}}], x, u)$ not contained in $\varphi([t + T_1, t + T_2], x, u)$ is contained in V_0. Hence each $a \in \varphi([t_{n_l}, t_{n_{l+1}}], x, u)$ has a distance less than $1/l$ to some $a' \in \varphi(T_1, T_2], y, v)$. Thus a second application of Lemma 4.4 yields

$$\varphi([T_1, T_2], y, v) = \partial \bigcap_l \operatorname{cl} \operatorname{ins} \Gamma_{n_l} = \partial \bigcap_n \operatorname{cl} \operatorname{ins} \Gamma_n = \omega(x, u).$$

Case 2 can be treated analogously.

The next theorem presents the main result of this paper.

THEOREM 4.6. *Let* $(x, u) \in \mathbb{R}^2_+ \times U_{\text{ad}}$ *be optimal for* (1.1)–(1.4) *with* $\varphi(\cdot, x, u)$ *nonself-intersecting and suppose that* $\omega(x, u)$ *contains only finitely many optimal equilibria. Then one of the following cases occurs*:

(i) *There are* $T > 0$ *and an optimal* $\mathbb{R}$*-solution* $(y, v) \in \hat{\omega}(x, u)$ *such that* $y = \varphi(T, y, v)$ *and* $\omega(x, u) = \varphi([0, T], y, v)$.

(ii) *There are optimal* $\mathbb{R}$*-solutions* $(y_i, v_i) \in \hat{\omega}(x, u)$ *and optimal equilibria* e_i^+, e_i^- *such that for all* i,

$$e_i^- = \lim_{t \to -\infty} \varphi(t, y_i, v_i), \qquad e_i^+ = \lim_{t \to +\infty} \varphi(t, y_i, v_i), \tag{4.1}$$

$$\omega(x, u) = \bigcup_i \varphi(\mathbb{R}, y_i, v_i) \cup \bigcup_i \{e_i^-, e_i^+\}. \tag{4.2}$$

Proof. Let $y \in \omega(x, u)$. By Proposition 2.10 there is $v \in U_{\text{ad}}$ such that $(y, v) \in \hat{\omega}(x, u)$. If either $\alpha(y, v)$ or $\omega(y, v)$ contain a point that is not an equilibrium, then Propositions 4.3 and 4.5 imply that (i) holds (naturally, we may take $T_1 = 0$).

In the other case, $\alpha(y, v)$ and $\omega(y, v)$ consist of equilibria only. Since $\alpha(y, v)$, $\omega(y, v) \subset \omega(x, u)$, Theorems 3.5 and 3.6 imply that there are optimal equilibria e^- and e^+ with

$$e^- = \lim_{t \to -\infty} \varphi(t, y, v), \qquad e^+ = \lim_{t \to +\infty} \varphi(t, y, v).$$

COROLLARY 4.7. *Let* $(x, u) \in \mathbb{R}^2_+ \times U_{\text{ad}}$ *be optimal for* (1.1)–(1.4) *and suppose that* $\omega(x, u)$ *does not contain an optimal equilibrium. Then either there is* $\hat{u} \in U_{\text{ad}}$ *such that* $(x, \hat{u})$ *is optimal, finally periodic and* $\varphi(x, \hat{u}) \subset \varphi(x, u)$ *or there are optimal periodic* $(y, v) \in \mathbb{R}^2_+ \times U_{\text{ad}}$ *with* $\omega(x, u) = \varphi(y, v)$.

Proof. If $\varphi(\cdot, x, u)$ is self-intersecting the assertion follows from Proposition 4.1. Otherwise Proposition 4.5 implies the existence of optimal $(\bar{y}, \bar{v})$ and $T_2 > T_1 \geqq 0$ with $\varphi(T_1, \bar{y}, \bar{v}) = \varphi(T_2, \bar{y}, \bar{v})$ and $\omega(x, u) = \varphi([T_1, T_2], y, v)$. Applying Proposition 4.1 again, we obtain the assertion.

Remark 4.8. By Remarks 2.12 and 3.7, the results above remain true if the performance criterion (1.1) is replaced by (2.5) provided that (2.6) holds.

5. Application to bioeconomic problems. The crucial assumption in the Poincaré-Bendixson Theorem given above is that the integrand of the performance criterion does not depend explicitly on the control u. In this section we show that the weakened form of this assumption specified in (2.6) can be verified in bioeconomic problems.

Furthermore, we present a specific example where the ω-limit set of an optimal solution consists of an optimal periodic trajectory that is not an optimal equilibrium. Feasibility of this case is a specific feature of the two-dimensional problem compared to the one-dimensional problem.

We will have to ensure that the ω-limit set has empty intersection with $\partial\mathbb{R}^2_+$. This deserves special attention also independently of the question considered here. Hence we give the following definition.

DEFINITION 5.1. A pair $(x, u) \in \operatorname{int}\mathbb{R}^n_+ \times U_{\mathrm{ad}}$ leads to extinction if $\omega(x, u) \cap \partial\mathbb{R}^n_+ \neq \emptyset$.

PROPOSITION 5.2. *Let* $(x, u) \in \operatorname{int}\mathbb{R}^n_+ \times U_{\mathrm{ad}}$ *be optimal. If* (x, u) *leads to extinction, then there are optimal* (y_k, v_k) $k = 0, 1, 2, \cdots$, *such that* $y_k \in \operatorname{int}\mathbb{R}^n_+$, $y_k \to y_0 \in \partial\mathbb{R}^n_+$ *and* $\sup\{d(z, \partial\mathbb{R}^n_+), z \in \varphi(y_k, v_k)\} \to 0$ *for* $k \to \infty$.

Proof. In the case $\omega(x, u) \subset \partial\mathbb{R}^n_+$ we may choose $y_k := \varphi(t_k, x, u)$, $v_k := u(t_k + \cdot)$, with $t_k \to \infty$. Now assume that $\omega(x, u) \cap \partial\mathbb{R}^n_+ \neq \emptyset$, but $\omega(x, u) \not\subset \partial\mathbb{R}^n_+$. For $\varepsilon > 0$ let $B_\varepsilon := \{z \in \mathbb{R}^n_+ : d(z, \partial\mathbb{R}^n_+) \leqq \varepsilon\}$. Choose ε small enough such that $\omega(x, u) \not\subset B_\varepsilon$. Then there are $t_l \to \infty$ and $s_l > 0$, $l \in \mathbb{N}$, such that for all l large enough

$$\varphi(t_l, x, u) \in \partial B_\varepsilon, \quad \varphi(t_l + s_l, x, u) \in B_{1/l}, \quad \varphi(t_l + s, x, u) \in B_\varepsilon \quad \text{for } s \in [0, s_l].$$

Without loss of generality we may assume that $t \to \varphi(t_l + t, x, u)$ converges locally uniformly to some $\varphi(\cdot, y, v)$ with $(y, v) \in \hat{\omega}(x, u)$, $y \in \partial B_\varepsilon$. If (s_l) is bounded we may assume that $s_l \to s \in \mathbb{R}_+$. This implies $\varphi(s, y, v) \in \partial\mathbb{R}^n_+$. Hence $\varphi(0, y, v) \subset \partial\mathbb{R}^n_+$. This contradicts $\varphi(t_l, x, u) \in \partial B_\varepsilon$. If (s_l) is unbounded, we may assume that $s_l \to \infty$. This implies $\varphi(y, v) \subset B_\varepsilon$. Choosing a sequence (ε_k) with $\varepsilon_k \to 0$, we obtain (y_k, v_k) satisfying the assertion.

Example 5.3. Maximize

$$V(a, u) = \int_0^\infty e^{-\delta t}\{p_1 x_1(\gamma_{11}u_1 + \gamma_{12}u_2)F^1(x_1) + p_2 x_2(\gamma_{21}u_1 + \gamma_{22}u_2)F^2(x_2) - c_1u_1 - c_2u_2\}\, dt$$

(where dependence on t has been dropped) such that

$$\begin{aligned}
&\dot{x} = x_1(F_0^1(x) - (\gamma_{11}u_1 + \gamma_{12}u_2)F^1(x_1)),\\
&\dot{x}_2 = x_2(F_0^2(x) - (\gamma_{21}u_1 + \gamma_{22}u_2)F^2(x_2)),\\
&(x_1(0), x_2(0)) = (a_1, a_2) \in \mathbb{R}^2_+,\\
&(u_1(t), u_2(t)) \in \Omega = [0, U_1] \times [0, U_2].
\end{aligned}$$

This example is designed to model resource-harvesting of two resources, the stocklevel of which (at time t) is denoted by $x_1(t)$, respectively, $x_2(t)$. There are two technologies available, such that an effort u_j spent applying technology j results in a catch-rate $\gamma_{ij}u_jF^i(x_i)$ with respect to the species i. γ_{ij} are nonnegative efficiency coefficients; $F^i(x_i)$ is a positive locally Lipschitz continuous function $\mathbb{R}^2_+ \to \mathbb{R}$, which relates effort and catch. There is a more detailed discussion of these "density profiles" in Clark [6].

p_1, p_2 are nonnegative constants to be interpreted as prices per unit biomass. c_1, c_2 are nonnegative constants to be interpreted as cost per unit effort spent applying technology one, respectively, two. Therefore $V(a, u)$ represents the "total discounted net revenue."

The rewriting of $V(a, u)$ *in* $\operatorname{int}\mathbb{R}^2_+$. Note first that

$$\gamma_{11}u_1+\gamma_{12}u_2=\frac{x_1F_0^1(x)-\dot{x}_1}{x_1F^1(x_1)},$$

$$\gamma_{21}u_1+\gamma_{22}u_2=\frac{x_2F_0^2(x)-\dot{x}_2}{x_2F^2(x_2)},$$

$$p_1x_1(\gamma_{11}u_1+\gamma_{12}u_2)F^1(x_1)=F_0^1(x)p_1x_1-p_1\dot{x}_1,$$

$$p_2x_2(\gamma_{21}u_1+\gamma_{22}u_2)F^2(x_2)=F_0^2(x)p_2x_2-p_2\dot{x}_2.$$

We assume that the matrix (γ_{ij}) is invertible. For obvious reasons the special case of $\gamma_{12}=\gamma_{21}=0$ is called "*selective harvesting.*" Suppose first $\gamma_{11}\neq 0$. Then

$$\left(\gamma_{22}-\frac{\gamma_{21}\gamma_{12}}{\gamma_{11}}\right)u_2=\frac{F_0^2(x)}{F^2(x_2)}-\frac{\gamma_{21}F_0^1(x)}{\gamma_{11}F^1(x_1)}-\frac{\dot{x}_2}{x_2F^2(x_2)}+\frac{\gamma_{21}}{\gamma_{11}}\frac{\dot{x}_1}{x_1F^1(x_1)},$$

or with $d=\gamma_{11}\gamma_{22}-\gamma_{12}\gamma_{21}$

$$u_2=G^2(x)-\frac{\gamma_{11}}{d}\frac{\dot{x}_2}{x_2F^2(x_2)}+\frac{\gamma_{21}}{d}\frac{\dot{x}_1}{x_1F^1(x_1)},$$

$$\begin{aligned}u_1&=G^1(x)+\frac{\gamma_{12}}{d}\frac{\dot{x}_2}{x_2F^2(x_2)}-\left(\frac{\gamma_{12}\gamma_{21}}{d\gamma_{11}}+\frac{1}{\gamma_{11}}\right)\frac{\dot{x}_1}{x_1F^1(x_1)}-c_1u_1-c_2u_2\\&=G^3(x)+\frac{\dot{x}_1}{x_1F^1(x_1)}\gamma_1+\frac{\dot{x}_2}{x_2F^2(x_2)}\gamma_2\end{aligned}$$

where

$$\gamma_1=\frac{c_1}{\gamma_{11}}+\frac{c_1\gamma_{12}\gamma_{21}}{d\gamma_{11}}-\frac{c_2\gamma_{21}}{d},\qquad \gamma_2=\frac{c_1\gamma_{12}}{d}-\frac{c_2\gamma_{11}}{d},$$

and G^1, G^2, G^3 are locally Lipschitz continuous functions of x. Put $G^4(x)=G^3(x)+F_0^1(x)p_1x_1+F_0^2(x)p_2x_2$. Then

$$V(a,u)=\int_0^\infty e^{-\delta t}G^4(x(t))\,dt+\int_0^\infty e^{-\delta t}\left\{\left[\frac{\gamma_1}{x_1F^1(x_1)}-p_1\right]\dot{x}_1+\left[\frac{\gamma_2}{x_2F^2(x_2)}-p_2\right]\dot{x}_2\right\}dt.$$

Put $g_j(y_j)=\int_{z_j}^{y_j}(\gamma_j/\xi F^j(\xi)-p_j)\,d\xi$, where $z=(z_1, z_2)$ is a pair of positive reals fixed once and for all. Now

$$\begin{aligned}V(a,u)&=\int_0^\infty e^{-\delta t}G^4(x(t))\,dt+\int_0^\infty e^{-\delta t}[g_1'(x_1(t))\dot{x}_1(t)+g_2'(x_2(t))\dot{x}_2(t)]\,dt\\&=\int_0^\infty e^{-\delta t}G^4(x(t))\,dt+\int_0^\infty e^{-\delta t}\frac{d}{dt}[g_1(x_1(t))+g_2(x_2(t))]\,dt\\&=\int_0^\infty e^{-\delta t}G^4(x(t))\,dt+e^{-\delta t}[g_1(x_1(t))+g_2(x_2(t))]\Big|_{t=0}^{\infty}\\&\quad+\int_0^\infty e^{-\delta t}\delta[g_1(x_1(t))+g_2(x_2(t))]\,dt\\&=r(a)+\int_0^\infty e^{-\delta t}R(x(t))\,dt\end{aligned}\tag{5.1}$$

with $r(a) := -g_1(a_1) - g_2(a_2)$, $R(y_1, y_2) = G^4(y) + \delta g_1(y_1) + \delta g_2(y_2)$, provided that

$$\lim_{t\to\infty} e^{\delta t}[g_1(x_1(t)) + g_2(x_2(t))] = 0. \tag{5.2}$$

To establish (5.2) let $c(t) := e^{-\delta t} g_j(x_j(t))$ for $j=1$ or $j=2$. It suffices to show that $\lim_{t\to\infty} c(t) = 0$. Now

$$\begin{aligned}\dot{c}(t) &= -\delta c(t) + e^{-\delta t}\left[\frac{\gamma_j}{x_j(t)F^j(x_j(t))} - p_j\right] x_j(t)[F_0^j(x(t)) - (y_{j1}u_1(t) + \gamma_{j2}u_2(t))F^j(x_j(t))] \\ &= -\delta c(t) + h(t),\end{aligned}$$

where h is a measurable and, since $F^j(0) \neq 0$, also bounded function on $\mathbb{R}_+$. By the variation of constant formula

$$|c(t)| \leqq (|c(0)| + \|h\|_\infty/\delta)\, e^{-\delta t}$$

and (5.2), and hence (5.1), follows.

The definition of R given above implies for $x_1, x_2 > 0$ that

$$\lim_{y\to(x_1,0)} R(y) = \pm\infty \quad \text{if } \gamma_2 \leqq 0,$$

$$\lim_{y\to(0,x_2)} R(y) = \pm\infty \quad \text{if } \gamma_1 \leqq 0.$$

Property (2.6)(α) follows from (5.1) and (5.2).

Analogous arguments can be used if $\gamma_{11} = 0$, and also in the case of nonselective harvesting, where $\gamma_{12} = \gamma_{21} = 0$.

Remark 5.4. We may construct examples of optimal control systems (with $m > n$) where condition (2.6) is not satisfied.

Now we present an example of a predator-prey system where both species are subject to innerspecific competition. Only the predator is harvested and the costs are proportional to the effort. The unharvested system possesses a limit cycle. We will show that there are optimal trajectories tending to an optimal periodic solution as $t \to \infty$. The system equation and the analysis of the uncontrolled system are taken from Sieveking [17].

Example 5.5.

$$\begin{aligned}\text{Maximize} \quad & \int_0^\infty e^{-\delta t}[pqx_2 - c]u\, dt \\ \text{Subject to} \quad & \dot{x}_1 = x_1[\alpha - \gamma x_2 - h(x_1) - \varepsilon x_1], \\ & \dot{x}_2 = x_2[-\beta + \lambda x_1 - \mu x_2 - qu], \qquad t \in \mathbb{R}_+, \\ & (x_1(0), x_2(0)) = x \in \mathbb{R}_+^2, \\ & u \in [0, U^1]\end{aligned}$$

where p, q, c, α, β, γ, δ, λ, μ, U^1 are positive constants, and h is defined by

$$h(x_1) = \begin{cases} (x_1 - \beta/\lambda)^2 & \text{for } 0 \leqq x_1 \leqq \beta/\lambda, \\ 0 & \text{for } \beta/\lambda \leqq x_1. \end{cases}$$

The system above is a special case of Example 5.3, and conditions (2.1), (2.6) are satisfied. First we analyze the uncontrolled equation where $u_1 = u_2 = 0$: All trajectories $\varphi(\cdot, x, 0)$ are bounded and for ε, $\mu > 0$, small, the only equilibria are $(0, 0)$, $(\alpha/\varepsilon, 0)$ and a point e near $e^0 = (\beta/\lambda, \alpha/\gamma)$.

The equilibrium e is (locally) asymptotically stable, the points $(0, 0)$ and $(\alpha/\varepsilon, 0)$ are saddles.

For ε, $\mu > 0$, small enough, the equation possesses a limit cycle (applying the Poincaré-Bendixson Theorem to the time-reversed equation, this implies the existence of another—unstable—periodic solution).

In the following, we assume that ε, μ are small enough such that existence of a limit cycle is guaranteed. There exists an initial value $x \in \operatorname{int} \mathbb{R}^2_+$ on the line $x_2 = -\beta/\mu + \lambda/\mu x_1$ such that $\varphi(\cdot, x, 0)$ spirals outward, i.e., there exists a (minimal) time $T_1 > 0$ such that $\varphi(T_1, x, 0)$ lies on the same line above x. Using continuous dependence of solutions on the right-hand side, this implies that for $U^1 > 0$, small enough, also every trajectory $\varphi(\cdot, x, u)$, $u(t) \in [0, U^1]$ almost everywhere, spirals outward. In particular, this is true for an optimal trajectory $\varphi(\cdot, x, u)$. See Fig. 5.1.

The controlled system has exactly two equilibria on $\partial \mathbb{R}^2_+$, namely $(0, 0)$ and $(\alpha/\varepsilon, 0)$.

Next we show that no optimal pair $(x, u) \in \operatorname{int} \mathbb{R}^2_+ \times U_{\text{ad}}$ leads to extinction. For $\xi > 0$, let $A_\xi := [\xi, \alpha/\varepsilon] \times [0, M]$ where $M := \max z_2(t)$ and $z = (z_1, z_2)$ is the unique trajectory in $\operatorname{int} \mathbb{R}^2_+$ of the uncontrolled system with $\lim_{t \to -\infty} z(t) = (\alpha/\varepsilon, 0)$. Then we can show that there exists $\xi > 0$ with the following property: For all $(y, v) \in \operatorname{int} \mathbb{R}^2_+ \times U_{\text{ad}}$ there is $T > 0$ such that for all $t \geqq T$ it follows that $\varphi(t, y, v) \in A_\xi$; furthermore $\varphi(y, v) \subset A_\xi$ for every $(y, v) \in A_\xi \times U_{\text{ad}}$. Hence $\omega(x, u) \cap \{0\} \times \mathbb{R}_+ = \varnothing$. Now suppose that $\omega(x, u) \cap \mathbb{R}_+ \times \{0\} \neq \varnothing$. Then Proposition 5.2 implies the existence of optimal $(y_k, v_k) \in \operatorname{int} \mathbb{R}^2_+ \times U_{\text{ad}}$ with $y_k \to y_0 \in \mathbb{R}_+ \times \{0\}$ and

$$\max \{d(z, \mathbb{R}_+ \times \{0\}) : z \in \varphi(y_k, v_k)\} \to 0 \quad \text{for } k \to \infty.$$

But for x_2 small, we have $pqx_2 - c < 0$. This contradicts the existence of (y_k, v_k) with the properties indicated above.

Conclusion. Suppose that in Example 5.5 the positive constants ε, μ, U^1 are small enough. No optimal pair $(x, u) \in \operatorname{int} \mathbb{R}^2_+ \times U_{\text{ad}}$ leads to extinction and every trajectory is bounded. There are initial values $x \in \operatorname{int} \mathbb{R}^2_+$ such that corresponding optimal trajectories $\varphi(\cdot, x, u)$ spiral outward. Hence, according to Corollary 4.7, there are optimal finally periodic $(x, \hat{u})$ or $\omega(x, u) = \varphi(y, v)$ with (y, v) optimal periodic.

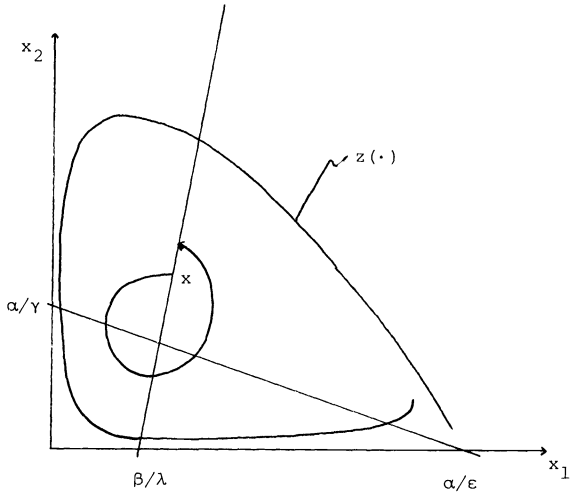

FIG. 5.1.

6. Nonuniqueness. For given initial state, solutions of ordinary differential equations are unique (provided that local Lipschitz continuity prevails). In general, optimal control problems do not share this nice property. In fact in this section we give a nonconstructive criterion implying nonuniqueness for a certain initial value. Furthermore, a simple bioeconomic example is presented with nonunique optimal solutions.

DEFINITION 6.1. An element $x \in \mathbb{R}^n_+$ is called a point of nonuniqueness if there are $u, v \in U_{\text{ad}}$ such that (x, u) and (x, v) are optimal and $\varphi(t, x, u) \neq \varphi(t, x, v)$ for some $t \in \mathbb{R}_+$.

"Nonuniqueness" requires that the trajectories corresponding to u and v do not coincide. Thus "redundancies" in the controls do not lead, in our terminology, to nonuniqueness.

THEOREM 6.2. *Suppose that* $(x, u) \in \mathbb{R}^2_+ \times U_{\text{ad}}$ *are optimal and that there are* $T_2 > T_1 \geqq 0$ *such that* $\varphi(\cdot, x, u)$, $t \in [T_1, T_2]$, *is a Jordan curve. If* $I := \text{cl ins } \Gamma$ *does not contain any optimal equilibrium, then it contains a point of nonuniqueness.*

Proof. Suppose there is no point of nonuniqueness in I and note that I is positively invariant. Hence for every $y \in I$, there is a unique control $u(y) \in U_{\text{ad}}$ such that $(y, u(y))$ is optimal and $\varphi(y, u(y)) \subset I$. Lemma 2.5 implies that $y \to u(y) : I \to U_{\text{ad}}$ is continuous, and hence for every $t \geqq 0$ the map $y \to \varphi(t, y, u(y)) : I \to I$ is continuous. By the Schoenfliess Theorem (Beck [3, p. 22]), I is homeomorphic to the closed unit ball in $\mathbb{R}^2$. Hence, by Brouwer's Fixed Point Theorem, there is for every $t \geqq 0$ a fixed point x_t with

$$\varphi(t, x_t, u(y)) = x_t.$$

Let (t_n) be a sequence of numbers with $t_n > 0$ such that $\lim t_n = 0$ and $\lim x_{t_n} = e \in I$ exists. We claim that e is an optimal equilibrium. In fact, for every $n \in \mathbb{N}$, uniqueness of optimal solutions implies that $\varphi(\cdot, x_n, u(x_n))$ is a periodic solution of period t_n. Without loss of generality we may assume that $\varphi(\cdot, x_n, u(x_n))$ converges uniformly to the constant trajectory e, which therefore is an optimal equilibrium contrary to our assumption.

In the following example, nonuniqueness is shown by a different argument.

Example 6.3.

$$\begin{aligned}
\text{Maximize} \quad & V(x, u) := \int_0^\infty e^{-\delta t}\{[p - c(x_1(t))]u_1(t)x_1(t) \\
& \qquad + [p - c(x_2(t))]u_2(t)x_2(t)\}\, dt \\
\text{Subject to} \quad & \dot{x}_1(t) = x_1(t)[2 - x_1(t) - 2x_2(t) - u_1(t)], \\
& \dot{x}_2(t) = x_2(t)[2 - x_2(t) - 2x_1(t) - u_2(t)], \\
& (u_1(t), u_2(t)) \in \Omega := [0, \tfrac{1}{2}] \times [0, \tfrac{1}{2}], \\
& x_1(0) = x_1, \qquad x_2(0) = x_2,
\end{aligned}$$

where $p > 0$ and $c(\cdot)$ is continuous and strictly decreasing on $\mathbb{R}_+$ with $c(\frac{2}{3}) = p$.

Assertion. For $\delta > 0$ sufficiently small the point $x = (\frac{2}{3}, \frac{2}{3})$ is a point of nonuniqueness.

Proof. (See Fig. 6.1.) First note that existence of an optimal solution follows by uniform boundedness of the trajectories, linearity in u and convexity and compactness of Ω. Define

$$S\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} y_2 \\ y_1 \end{pmatrix}, \qquad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \in \mathbb{R}^2_+.$$

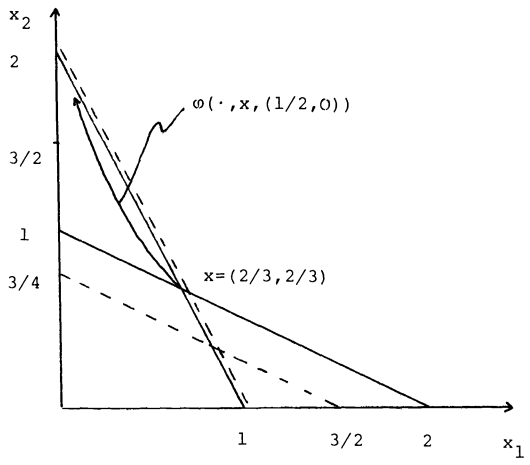

FIG. 6.1. *Illustration of Example* 6.3.

The symmetry of the system equation and $x = Sx$ imply for $u \in U_{ad}$ and $t \geqq 0$

$$S\varphi(t, x, u) = \varphi(t, x, Su).$$

Furthermore,

$$V(x, Su) = V(x, u).$$

Thus, if (x, u) is optimal, also (x, Su) is optimal. If the optimal solution is unique, it follows that

$$\varphi_1(t, x, u) = \varphi_2(t, x, u) \quad \text{for all } t \geqq 0.$$

Looking at the system equation we find that this implies

$$\varphi_1(t, x, u) \leqq \tfrac{2}{3}, \quad \varphi_2(t, x, u) \leqq \tfrac{2}{3} \quad \text{for all } t > 0.$$

Hence

$$p - c(\varphi_1(t, x, u)) \leqq 0, \quad p - c(\varphi_2(t, x, u)) \leqq 0 \quad \text{for all } t > 0$$

and

$$V(x, u) \leqq 0.$$

Thus, in case of uniqueness, the only candidate for an optimal control is $u_1 \equiv u_2 \equiv 0$, which leaves $x = (\frac{2}{3}, \frac{2}{3})$ fixed and

$$V(x, u) = 0.$$

Thus it suffices to construct $v \in U_{ad}$ with

$$V(x, v) > 0.$$

Consider first the control $\bar{v} = (\bar{v}_1, \bar{v}_2)$

$$\bar{v}_1(t) \equiv \tfrac{1}{2}, \qquad \bar{v}_2(t) \equiv 0.$$

A phase plane analysis (cf. Fig. 6.1) yields that for t increasing, $\varphi_1(t, x, \bar{v})$ decreases and $\varphi_2(t, x, \bar{v})$ increases, with

$$\lim_{t \to \infty} \varphi(t, x, \bar{v}) = (0, 2). \tag{6.1}$$

Now consider the system

$$(6.2)\qquad \dot{x}_1(t) = x_1(t)[2 - x_1(t) - 2x_2(t)], \qquad \dot{x}_2(t) = x_2(t)[2 - x_2(t) - 2x_1(t) - \tfrac{1}{2}].$$

For this system the point $(0, \frac{3}{2})$ is (locally) asymptotically stable. In fact, the Jacobian at this point is

$$\begin{pmatrix} -1 & 0 \\ -3 & -\frac{3}{2} \end{pmatrix}.$$

Since the region of attraction of an asymptotically stable point is always open and $(0, 2)$ is attracted by $(0, \frac{3}{2})$, it follows from (6.1) that there is $t_1 > 0$ such that in the system (6.2), $\varphi(t_1, x, \bar{v})$ is attracted by $(0, \frac{3}{2})$. Define

$$v(t) := \begin{cases} \bar{v}(t) = (\frac{1}{2}, 0), & t \in [0, t_1], \\ (0, \frac{1}{2}), & t \in (t_1, \infty). \end{cases}$$

Then

$$\lim_{t \to \infty} \varphi(t, x, v) = (0, \tfrac{3}{2}).$$

By continuity of c, there is $M_1 > 0$ with

$$(p_1 - c(\varphi_1(t, x, \bar{v})))\varphi_1(t, x, \bar{v})\tfrac{1}{2} \geqq -M_1, \qquad 0 \leqq t \leqq t_1.$$

Thus

$$\begin{aligned} \delta V(x, v) &= \delta \int_0^\infty e^{-\delta t}\{[p - c(\varphi_1(t, x, v))]\varphi_1(t, x, v)v_1(t) \\ &\qquad + [p - c(\varphi_2(t, x, v))]\varphi_2(t, x, v)v_2(t)\}\, dt \\ &= \delta \int_0^{t_1} \cdots + \delta \int_{t_1}^\infty \cdots \\ &\geqq -\delta M_1 \frac{1}{2} \int_0^{t_1} e^{-\delta t}\, dt + \delta \int_{t_1}^\infty e^{-\delta t}[p - c(\varphi_2(t, x, \bar{v}))]\frac{1}{2}\, dt; \end{aligned}$$

without loss of generality we may assume

$$\varphi_2(t, x, v) \geqq 1 \quad \text{for all } t \geqq t_1.$$

Since there is $M_2 > 0$ with

$$p - c(y) > M_2 \quad \text{for } y \geqq 1$$

it follows that

$$(p - c(\varphi_2(t, x, v)))\varphi_2(t, x, v)\tfrac{1}{2} \geqq \tfrac{1}{2}M_2 \quad \text{for } t \geqq t_1.$$

Together we get

$$V(x, v) \geqq -\tfrac{1}{2}M_1(1 - e^{-\delta t_1}) + \tfrac{1}{2}M_2\, e^{-\delta t_1}.$$

For $\delta \to 0$, the right-hand side of this inequality tends to $\frac{1}{2}M_2$. Hence $V(x, \bar{v}) > 0$ for $\delta > 0$, sufficiently small. This proves the assertion.

The idea for this example may be sketched as follows. We start in an equilibrium point x, where two competing species coexist, and where the net revenue $p - c(x)$ is zero. Catching *one* of these species we have a temporary loss. On the other hand, the other species increases until it gets into a domain where it can be caught continually, yielding positive net revenue. The initial loss is, for sufficiently small discount rate $\delta > 0$, less than the later revenue.

Acknowledgments. The work of C. W. Clark as well as conversations with him have been a source of motivation and ideas for the present paper. Furthermore, we thank D. Hinrichsen for the support that he provided to our work.

REFERENCES

[1] H. AMANN, *Gewöhnliche Differentialgleichungen*, De Gruyter, Berlin, 1983.
[2] K. J. ARROW, *Applications of control theory to economic growth*, in Mathematics of the Decision Sciences, G. B. Dantzig and A. F. Veinott, eds., American Mathematical Society, Providence, RI, 1968, pp. 85-119.
[3] A. BECK, *Continuous Flows in the Plane*, Springer-Verlag, Berlin, New York, 1974.
[4] J. BENHABIB AND K. NISHIMURA, *The Hopf bifurcation and the existence and stability of closed orbits in multisector models of optimal economic growth*, J. Econom. Theory, 21 (1979), pp. 421-444.
[5] C. W. CLARK, *Mathematical Bioeconomics: The Optimal Management of Renewable Resources*, John Wiley, New York, 1976.
[6] ———, *Bioeconomic Modelling and Fisheries Management*, John Wiley, New York, 1985.
[7] P. DEKLERK AND M. GATTO, *Some remarks on periodic harvesting of a fish population*, Math. Biosci., 56 (1981), pp. 47-69.
[8] C. D. FEINSTEIN AND D. G. LUENBERGER, *Characterization of the asymptotic behaviour of optimal control trajectories: The implicit programming problem*, SIAM J. Control Optim., 19 (1981), pp. 561-585.
[9] C. D. FEINSTEIN AND S. S. OREN, *A "funnel" turnpike theorem for optimal growth problems with discounting*, J. Econom. Dyn. Control, 9 (1985), pp. 25-39.
[10] O. HAJEK, *Dynamical Systems in the Plane*, Academic Press, New York, 1968.
[11] H. HALKIN, *Necessary conditions for optimal control problems with infinite horizons*, Econometrica, 42 (1974), pp. 267-272.
[12] A. HAURIE, *Existence and global asymptotic stability of optimal trajectories for a class of infinite horizon nonconvex systems*, J. Optim. Theory Appl., 31 (1980), pp. 515-533.
[13] ———, *Stability and optimal exploitation over an infinite horizon of interacting populations*, Optimal Control Appl. Methods, 3 (1982), pp. 241-256.
[14] M. W. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems and Linear Algebra*, Academic Press, New York, 1974.
[15] R. T. ROCKAFELLAR, *Saddle points of Hamiltonian systems in convex problems of Lagrange*, in The Hamiltonian Approach to Dynamic Economics, D. Cass and K. Shell, eds., Academic Press, New York, 1976.
[16] ———, *Saddle points of Hamiltonian systems in convex Lagrange problems having a nonzero discount rate*, J. Econom. Theory, 12 (1976), pp. 71-113.
[17] M. SIEVEKING, *Dinamica de poblaciones por medio de ecuaciones differenciales ordinarias*, Escuela Politecnica Nacional, Quito, Ecuador, and Fb. Mathematik der J. W. Goethe, Universität Frankfurt, Frankfurt, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 631-641, May 1989

010

ASYMPTOTIC ADMISSIBILITY OF THE UNIT STEPSIZE IN EXACT PENALTY METHODS*

JOSEPH FRÉDÉRIC BONNANS†

Abstract. Two difficulties arise in the use of optimization algorithms based on an exact penalty function and quadratic subproblems: the possible inconsistency of the quadratic programs and the admissibility of the unit stepsize after a finite number of iterations. In this paper, assuming that no inequality constraint is present, the author devises an algorithm, using a nondifferentiable augmented Lagrangian, that, under convenient hypotheses, solves both problems.

Key words. nonlinear optimization, constrained optimization, Newton's method, successive quadratic programming, optimization algorithms

AMS(MOS) subject classifications. 90C30, 65K05, 49D15

1. Introduction.

We consider a nonlinear programming problem having only equality constraints:

(1.1) Minimize $f(x)$ subject to $g_i(x)=0$, $i=1$ to m,

f and g_i, $i=1$ to m being smooth (C^3) functions from $\mathbb{R}^n$ into $\mathbb{R}$. We suppose that $m \leqq n$. Let $\bar{x}$ be a local solution of (1.1). We suppose that $\bar{x}$ is a regular point, i.e.,

(1.2) $\nabla g_i(\bar{x})$, $i=1$ to m, are linearly independent.

Then there exists a unique $\bar{\lambda} \in \mathbb{R}^m$ satisfying

$$\nabla f(\bar{x}) + \nabla g(\bar{x})\bar{\lambda} = 0, \qquad g(\bar{x}) = 0. \tag{1.3}$$

Equations (1.3) may be solved by a Newton-type method with unknowns (x, λ). This reduces to the computation of a sequence (x^k, λ^k) with $x^{k+1} = x^k + d^k$, where d^k is a solution of the quadratic program

(1.4) Minimize $\nabla f(x^k)'d + \frac{1}{2}d'H^k d$ subject to $g(x^k) + \nabla g(x^k)'d = 0$,

H^k being an approximation to the Hessian of the Lagrangian, and λ^{k+1} being the multiplier associated with d^k. We restrict our analysis to the case when H^k is a positive definite approximation of the Hessian of the Lagrangian. When this approximation is made using the BFGS (Broyden-Fletcher-Goldfarb-Shanno) method, a superlinear convergence rate may be obtained (Powell [15], Boggs, Tolle, and Wang [2]). To globalize this algorithm—i.e., to design some globally convergent algorithm that reduces to a Newton-type method in the neighbourhood of a solution—a key idea is to use the nondifferentiable penalty function

$$\theta_r(x) = f(x) + r\|g(x)\|,$$

where $r>0$ is called the penalty parameter and $\|\cdot\|$ is some norm of $\mathbb{R}^m$; we denote by $\|\cdot\|_D$ the dual norm, i.e.,

$$\|\lambda\|_D = \max\left\{ \sum_{i=1}^{m} \lambda_i \mu_i,\ \|\mu\| = 1 \right\}.$$

* Received by the editors July 6, 1987; accepted for publication (in revised form) September 15, 1988.
† Institut National de Recherche en Informatique et Automatique, Rocquencourt, 78153 Le Chesnay, France.

A key fact is that if (1.4) has a solution d^k associated to a multiplier λ^{k+1}, and if $r > \|\lambda^{k+1}\|_D$, then d^k is a descent direction of $\theta_r(x)$ at x^k. This result is due to Pschenichnyi and Danilin (see [17]) in the case of the L^∞ norm and was rediscovered by Han [10] in the case of the L^1 norm. The general result can be found in Bonnans and Gabay [4]. A difficulty arises if the m vectors $\{\nabla g_i(x^k)\}$ are not everywhere linearly independent; then (1.4) may have no solutions. Some empirical means for taking this into account are given in Powell [14] and Tone [18]. The algorithm of Bonnans and Gabay [4] seems, at least from a theoretical point of view, to give a satisfactory solution to this problem. It is based on the function, defined for each iteration k:

$$\theta_r^k(x) = f(x^k) + \nabla f(x^k)'(x - x^k) + \tfrac{1}{2}(x - x^k)' H^k (x - x^k) + r\|g(x^k) + \nabla g(x^k)'(x - x^k)\|,$$

which is a simple model of $\theta_r(x)$, having the same behaviour around x^k. A direction $\tilde{d}^k$ is computed as the solution of

$$\min\{\theta_r^k(x^k + d);\ d \in \mathbb{R}^n\}. \tag{1.5}$$

Being strictly convex and feasible, problem (1.5) always has a unique solution $\tilde{d}^k$. Then the iteration is

$$x^{k+1} = x^k + \rho^k \tilde{d}^k,$$

where ρ^k is a stepsize computed according to some linesearch rule. Define

$$\tilde{\theta}_r^k(x) = f(x^k) + \nabla f(x^k)'(x - x^k) + r\|g(x^k) + \nabla g(x^k)'(x - x^k)\|.$$

A convenient linesearch rule (see, e.g., Chamberlain et al. [6]) is the following extension of the Armijo rule [1]:

(1.6) Choose $\beta \in]0, 1[$, $\sigma \in]0, \frac{1}{2}[$ (independent of k),
$\rho^k = (\beta)^l$, where l is the smallest nonnegative integer such that
$\theta_r(x^k + (\beta)^l \tilde{d}^k) - \theta_r(x^k) \leqq \sigma(\tilde{\theta}_r^k(x^k + (\beta)^l \tilde{d}^k) - \theta_r(x^k))$.

This means that we reduce the step until the ratio of the achieved decrease on the penalty function divided by the decrease predicted by the local model $\tilde{\theta}_r^k$ is at least σ.

The complexity of problem (1.5) is, at least for the L^1 and L^∞ norms, roughly the same as that of a quadratic program. In addition, if (1.4) has a solution d^k associated with a unique multiplier λ^{k+1}, the solution of (1.5) will be equal to d^k if r^k is greater than $\|\lambda^{k+1}\|_D$. Consequently, when the parameter r^k is iteratively modified in a convenient way, the method leads to a globally convergent algorithm, where computed displacements reduce to the solution of (1.4) near a regular solution of (1.1). This method is related to that of Fletcher [7], [8] who uses a trust region method instead of a linesearch.

One point should be clarified. To compute descent directions, the algorithm described above relaxes the linearized constraints by penalizing them. This is useful when these linearized constraints are not necessarily compatible at any point. However, if it is known a priori that the linearization of some subset of the constraints are compatible, there is no reason to relax the constraints of this subset. A key fact is that this subset must contain the linear constraints (otherwise, the problem has no solution). This means that, in practice, the linear constraints should not be relaxed. To keep the proofs short, in this paper we do not take this remark into account. However, this modification of the algorithm should not essentially change the results.

We now turn to the local convergence analysis. We suppose that a sequence $\{x^k\}$ computed by the preceding algorithm converges to a local solution $\bar{x}$ of (1.1), which is regular point. Let d^k be the solution of (1.4) (well defined for k large enough). Let $\|\cdot\|_U$ be a norm of $\mathbb{R}^n$. We suppose that the vector d^k, which solves (1.4), satisfies

$$\|x^k + d^k - \bar{x}\|_U / \|x^k - \bar{x}\|_U \to 0. \tag{1.7}$$

This hypothesis allows the algorithm to have a superlinear rate of convergence, if $\{\rho^k\}$ converges to 1. However, this does not seem to be the case in general: a counterexample due to Maratos [12] shows that even if (1.7) holds, and if r has the same order of magnitude as $\|\bar{\lambda}\|_D$, $\theta_r(x^k + \rho^k d^k)$ may be greater than $\theta_r(x^k)$ when $\rho^k \to 1$. This may destroy the property of superlinear convergence. Two kind of methods have been proposed to deal with this problem. The first (Mayne and Polak [13]) needs the computation of the constraints at $x^k + d^k$; then a correction term v^k is computed as the solution of

$$\text{Minimize } \sum_{i=1}^{n} (v_i^k)^2 \quad \text{subject to } g(x^k + d^k) + \nabla g(x^k)^t v^k = 0.$$

If a linesearch is used, it can be performed along the arc

$$x^k + \rho d^k + (\rho)^2 v^k.$$

Then it is shown that, under some convenient assumptions, the stepsize $\rho^k = 1$ is admissible if k is great enough. The second method, due to Chamberlain et al. [6], is based on the observation that a sufficient decrease of the exact penalty function is obtained (for k large enough) between the iterations $k-1$ and $k+1$. Consequently, the linesearch criterion at step k should use the information of the iteration $k-1$. However, if the point $x^k + d^k$ is not accepted, we must return to the point x^{k-1} and reduce the stepsize at x^{k-1}; this ensures global convergence. The main drawback of these methods is that, in some situations, they can substantially increase the amount of computations.

In this paper we propose to perform the linesearch, using the following criterion:

$$\theta_{p,r}(x) = f(x) + p^t g(x) + r\|g(x)\|.$$

Here p is an approximation of the optimal Lagrange multiplier $\bar{\lambda}$ and $r > 0$ is a penalty parameter, as before. Choosing p close to $\bar{\lambda}$ allows us to reduce the value of r; we will prove that, if p and r are carefully adapted at each iteration, r being small enough, the unit stepsize is asymptotically admissible. This result allows us to build a "globally convergent" algorithm.

The paper is organized as follows. In § 2 we give a technical result on exact penalty functions that is the basis of the subsequent algorithm. Then in § 3 we use this result to formulate a globally convergent method, based on a linesearch strategy, for which the unit stepsize is, under some convenient hypothesis, asymptotically admissible.

2. Some local properties of a class of exact penalty functions. Let $\bar{x}$ be a local solution of (1.1) satisfying (1.2) and $\bar{\lambda}$ be the element of $\mathbb{R}$ such that (1.3) holds. Define the augmented Lagrangian (for $c \geqq 0$)

$$L_c(x, \lambda) = f(x) + \lambda^t g(x) + c \sum_{i=1}^{m} g_i(x)^2.$$

Denote

$$H_c = \frac{\partial^2 L_c(\bar{x}, \bar{\lambda})}{\partial x^2}.$$

We suppose that the standard second-order sufficiency condition holds at $\bar{x}$ (see, for instance, Fletcher [7]):

(2.1) $d^t H_0 d > 0$ for any d in $\mathbb{R}^n$ satisfying $\nabla g(x)^t d = 0$.

As is well known, (1.3) and (2.1) imply that

(2.2) There exists $\bar{c} > 0$ such that $H_{\bar{c}}$ is positive definite.

We now consider the following class of penalty function that can be viewed as a class of nondifferentiable augmented Lagrangians:

$$\theta_{p,r}(x) = f(x) + p^t g(x) + r\|g(x)\|,$$

where $(p, r) \in \mathbb{R}^m \times \mathbb{R}$ are given parameters with $r > 0$. We recall that $\|\cdot\|$ is a norm of $\mathbb{R}^m$ and $\|\cdot\|_D$ is its dual norm. We give a sufficient condition for these penalty functions to be exact, i.e., to have a (strict) local minimum at $\bar{x}$. This is a variant of results of Han and Mangasarian [11] (see also Bonnans [3]).

PROPOSITION 2.1. *Let $\bar{x}$ be a local minimum of* (1.1) *such that, for some* $\bar{\lambda}$, (1.3) *and* (2.1) *hold. Then if*

(2.3) $$r > \|\bar{\lambda} - p\|_D,$$

$\bar{x}$ is a strict local minimum of $\theta_{p,r}(x)$.

Remark 2.1. As all norms on $\mathbb{R}^m$ are equivalent, there exists $\beta > 0$ satisfying

$$\sum_{i=1}^{m} g_i(x)^2 \leqq \beta \|g(x)\|^2.$$

Proof of Proposition 2.1. We prove that the penalty function $\theta_{p,r}$ dominates the augmented Lagrangian. We have

$$\begin{aligned}\theta_{p,r}(x) - L_{\bar{c}}(x, \bar{\lambda}) &= (p - \bar{\lambda})^t g(x) + r\|g(x)\| - \bar{c} \sum_{i=1}^{m} g_i(x)^2, \\ &\geqq (r - \|p - \bar{\lambda}\|_D)\|g(x)\| \bar{c} \sum_{i=1}^{m} g_i(x)^2, \\ &\geqq (r - \|p - \bar{\lambda}\|_D - \bar{c}\beta\|g(x)\|)\|g(x)\|.\end{aligned}$$

As $g(\bar{x}) = 0$, if (2.3) holds, the right-hand side is nonnegative in a neighbourhood of $\bar{x}$. As $H_{\bar{c}}$ is positive definite, $\bar{x}$ is a strict local minimum of $L_{\bar{c}}(x, \bar{\lambda})$; this proves the proposition. □

Let $\{x^k\}$ be a sequence converging to $\bar{x}$ and $\{p^k\}, \{r^k\}$ be two sequences such that, for k large enough and for some $\gamma > 0$,

(2.4) $$(3 + \gamma)\|p^k - \bar{\lambda}\|_D < r^k,$$

(2.5) $$r^k / \|x^k - \bar{x}\|_U \to +\infty.$$

We define

$$\begin{aligned}\tilde{\theta}^k(x) = f(x^k) + \nabla f(x^k)^t (x - x^k) + (p^k)^t (g(x^k) + \nabla g(x^k)^t (x - x^k)) \\ + r^k \|g(x^k) + \nabla g(x^k)^t (x - x^k)\|.\end{aligned}$$

We recall that the solution d^k of (1.4) is well defined in a neighbourhood of $\bar{x}$ if H^k is positive definite.

THEOREM 2.1. *We suppose that* (1.3), (1.7), (2.1), (2.4), *and* (2.5) *hold. Then for any* ϵ, $0<\varepsilon<\frac{1}{2}$, *there exists* $r_0>0$ *and* k_0 *such that* $k>k_0$ *and* $0<r^k<r_0$ *imply*

$$\theta_{p^k,r^k}(x^k+d^k)-\theta_{p^k,r^k}(x^k)\leqq(\tfrac{1}{2}-\varepsilon)(\tilde{\theta}^k(x^k+d^k)-\theta_{p^k,r^k}(x^k)). \tag{2.6}$$

The proof of the theorem uses two lemmas.

LEMMA 2.1. *We suppose that* (1.3), (1.7), *and* (2.1) *hold. Then, for any* $\varepsilon_1>0$, *there exists* k_1 *such that* $k>k_1$ *implies*

$$\tilde{\theta}^k(x^k+d^k)-\theta_{p^k,r^k}(x^k)\geqq-(1+\varepsilon_1)(x^k-\bar{x})^tH_{\bar{c}}(x^k-\bar{x})-(r^k+\|p^k-\bar{\lambda}\|_D)\|g(x^k)\|.$$

Proof. We have for k large enough

$$\begin{aligned}\Delta&=\tilde{\theta}^k(x^k+d^k)-\theta_{p^k,r^k}(x^k)=\nabla f(x^k)^td^k+(p^k)^t\nabla g(x^k)^td^k-r^k\|g(x^k)\|,\\&=\nabla_xL_0(x^k,\bar{\lambda})^td^k+(p^k-\bar{\lambda})^t\nabla g(x^k)^td^k-r^k\|g(x^k)\|.\end{aligned}$$

From (1.4) it follows that

$$\begin{aligned}\Delta&=\nabla_xL_0(x^k,\bar{\lambda})^td^k-(p^k-\bar{\lambda})^tg(x^k)-r^k\|g(x^k)\|,\\&\geqq\nabla_xL_0(x^k,\bar{\lambda})^td^k-(r^k+\|p^k-\bar{\lambda}\|_D)\|g(x^k)\|.\end{aligned}$$

Now (1.3) and (1.7) imply

$$\begin{aligned}\nabla_xL_0(x^k,\bar{\lambda})&=H_0(x^k-\bar{x})+o(\|x^k-\bar{x}\|_U),\\d^k&=\bar{x}-x^k+o(\|x^k-\bar{x}\|_U),\end{aligned}$$

where the notation $o(\|x^k-\bar{x}\|_U)$ indicates a term whose ratio to $\|x^k-\bar{x}\|_U$ tends to zero as $k\to\infty$. We deduce that

$$\Delta\geqq-(x^k-\bar{x})H_0(x^k-\bar{x})-(r^k+\|p^k-\bar{\lambda}\|_D)\|g(x^k)\|+o(\|x^k-\bar{x}\|_U^2).$$

The result is then a consequence of the inequality

$$d^tH_{\bar{c}}d\geqq d^tH_0d\quad\forall d\in\mathbb{R}^n,$$

and of the positive definiteness of $H_{\bar{c}}$. □

LEMMA 2.2. *If* (1.3), (1.7), *and* (2.1) *hold, for any* $\varepsilon_2>0$, *there exists* k_2 *such that* $k>k_2$ *implies*

$$\begin{aligned}\theta_{p^k,r^k}(x^k+d^k)-\theta_{p^k,r^k}(x^k)\leqq&-\tfrac{1}{2}(1-\varepsilon_2)(x^k-\bar{x})^tH_{\bar{c}}(x^k-\bar{x})\\&+(\|p^k-\bar{\lambda}\|_D+r^k)\|g(x^k+d^k)\|\\&-(r^k-\|p^k-\bar{\lambda}\|_D-\bar{c}\beta\|g(x^k)\|)\|g(x^k)\|.\end{aligned} \tag{2.7}$$

Proof. We have

$$\begin{aligned}\Delta&=\theta_{p^k,r^k}(x^k+d^k)-\theta_{p^k,r^k}(x^k),\\&=L_{\bar{c}}(x^k+d^k)-L_{\bar{c}}(x^k)+(p^k-\bar{\lambda})^t(g(x^k+d^k)-g(x^k))\\&\quad-\bar{c}\left(\sum_{i=1}^m g_i(x^k+d^k)^2-\sum_{i=1}^m g_i(x^k)^2\right)+r^k(\|g(x^k+d^k)\|-\|g(x^k)\|).\end{aligned}$$

Hence, using Remark 2.1, we have

$$\begin{aligned}\Delta\leqq&L_{\bar{c}}(x^k+d^k)-L_{\bar{c}}(x^k)+(\|p^k-\bar{\lambda}\|_D+r^k)\|g(x^k+d^k)\|\\&-(r^k-\|p^k-\bar{\lambda}\|_D-\bar{c}\beta\|g(x^k)\|)\|g(x^k)\|.\end{aligned} \tag{2.8}$$

We focus on the first terms, using (1.3) and (1.7):

$$L_{\bar c}(x^k+d^k)-L_{\bar c}(x^k)=L_{\bar c}(x^k+d^k)-L_{\bar c}(\bar x)+L_{\bar c}(\bar x)-L_{\bar c}(x^k)$$
$$=-\tfrac{1}{2}(x^k-\bar x)^t H_{\bar c}(x^k-\bar x)+o(\|x^k-\bar x\|_U^2).$$

This with (2.8) and the positive definiteness of $H_{\bar c}$ imply (2.7). □

Proof of Theorem 2.1. From the constraints of (1.4) and (1.7) we deduce the existence of some $a_1>0$ such that for k large enough

$$\|g(x^k+d^k)\|\leqq a_1\|x^k-\bar x\|_U^2.$$

From Lemma 2.2 and the positive definiteness of $H_{\bar c}$, it follows that for some $a_2>0$ and k large enough

$$\Delta_1=\theta_{p^k,r^k}(x^k+d^k)-\theta_{p^k,r^k}(x^k)$$
$$\leqq-\tfrac{1}{2}[1-\varepsilon_2-a_2(\|p^k-\bar\lambda\|_D+r^k)](x^k-\bar x)^t H_{\bar c}(x^k-\bar x)$$
$$-(r^k-\|p^k-\bar\lambda\|_D-\bar c\beta\|g(x^k)\|)\|g(x^k)\|.$$

Using (2.4) and (2.5) we deduce that for k large enough, and r_0 small enough,

$$\Delta_1\leqq-\tfrac{1}{2}(1-2\varepsilon_2)(x^k-\bar x)^t H_{\bar c}(x^k-\bar x)-\tfrac{2}{3}r^k\|g(x^k)\|.$$

From Lemma 2.1 and (2.4) we deduce that for k large enough

$$\Delta_2=\tilde\theta^k(x^k+d^k)-\theta_{p^k,r^k}(x^k)\geqq-(1+\varepsilon_1)(x^k-\bar x)^t H_{\bar c}(x^k-\bar x)-\tfrac{4}{3}r^k\|g(x^k)\|.$$

We must prove that $(\frac{1}{2}-\varepsilon)\Delta_2-\Delta_1\geqq 0$. Using the inequalities above, we find (for $0<\varepsilon<\frac{1}{2}$ and k large enough)

$$(\tfrac{1}{2}-\varepsilon)\Delta_2-\Delta_1\geqq\tfrac{1}{2}[1-2\varepsilon_2-(1-2\varepsilon)(1+\varepsilon_1)](x^k-\bar x)^t H_{\bar c}(x^k-\bar x)+\tfrac{4}{3}\varepsilon r^k\|g(x^k)\|,$$
$$\geqq\tfrac{1}{2}(2\varepsilon-2\varepsilon_2-\varepsilon_1)(x^k-\bar x)^t H_{\bar c}(x^k-\bar x).$$

Since ε_1 and ε_2 may be taken arbitrarily small for r_0 small enough and k large enough, we get the result. □

This result suggests building an algorithm, using a penalty function of type $\theta_{p^k,r^k}(x)$, where p^k and r^k are modified at each iteration to ensure a global convergence, which satisfies the hypothesis of Theorem 2.1 after a finite number of iterations. This is the subject of the following section.

3. A globally and superlinearly convergent algorithm. We define a kind of quadratic model of θ_{p^k,r^k} around x^k:

$$\theta^k(x)=\tilde\theta^k(x)+\tfrac{1}{2}(x-x^k)^t H^k(x-x^k). \tag{3.1}$$

We consider the following algorithm.

ALGORITHM 1.

(0) Choose x^1, p^1, r^1, H^1 such that $r^1>0$ and H^1 is positive definite. Set $k=1$. Choose $\beta\in]0,1[$, $\sigma\in]0,\frac{1}{2}[$.

(1) Solve the problem

$$\min\ \theta^k(x^k+d),\qquad d\in\mathbb{R}^n. \tag{3.2}$$

Let $\tilde d^k$ be the unique solution of (3.2). If $\tilde d^k=0$, stop.

(2) Let l be the smallest nonnegative integer such that

$$\theta_{p^k,r^k}(x^k+(\beta)^l\tilde d^k)-\theta_{p^k,r^k}(x^k)\leqq\sigma(\tilde\theta^k(x^k+(\beta)^l\tilde d^k)-\theta_{p^k,r^k}(x^k)), \tag{3.3}$$
$$\rho^k=(\beta)^l,\qquad x^{k+1}=x^k+\rho^k\tilde d^k.$$

(3) $k=k+1$. Set p^k, r^k, H^k with $r^k>0$ and H^k positive definite. Go to (1).

We remark that θ_{p^k,r^k} and $\tilde{\theta}^k$ have the same directional derivatives at x^k and that this derivative in the direction $\tilde{d}^k$ is strictly negative; hence, the integer l of step (2) is well defined.

We now proceed to give an explicit adaptation rule for p^k and r^k. This needs some preliminary considerations. Consider

$$Q^k(\lambda) = \tfrac{1}{2}\lambda^t \nabla g(x^k)^t (H^k)^{-1} \nabla g(x^k)\lambda + \lambda^t (\nabla g(x^k)^t (H^k)^{-1} \nabla f(x^k) - g(x^k)).$$

It is well known that the quadratic program (1.4) is equivalent to

$$\text{(3.4)} \quad \begin{array}{ll} \text{(i)} & \lambda^{k+1} = \arg\min Q^k(\lambda), \qquad \lambda \in \mathbb{R}^m, \\ \text{(ii)} & d^k = -(H^k)^{-1}(\nabla f(x^k) + \nabla g(x^k)\lambda^{k+1}). \end{array}$$

On the other hand, from Bonnans and Gabay [4] we deduce that (3.2) is equivalent to

$$\text{(3.5)} \quad \begin{aligned} &\tilde{\lambda}^{k+1} = \arg\min \{Q^k(\lambda), \|\lambda - p^k\|_D \leqq r^k\}, \\ &\tilde{d}^k = -(H^k)^{-1}(\nabla f(x^k) + \nabla g(x^k)\tilde{\lambda}^{k+1}). \end{aligned}$$

We deduce from (3.4) and (3.5) that

$$\text{(3.6)} \quad \{\|\tilde{\lambda}^{k+1} - p^k\|_D < r^k \Rightarrow \{\text{(1.4) has a solution } (d^k, \lambda^{k+1}) \text{ equal to } (\tilde{d}^k, \tilde{\lambda}^{k+1})\}.$$

We define the sequences

$$D^k = \|\nabla f(x^k) + \nabla g(x^k)\tilde{\lambda}^k\|_U + \|g(x^k)\|,$$

$$S^k = \max\{1/D^l, l = 1 \text{ to } k\}.$$

The monotonically nondecreasing sequence $\{S^k\}$ has the following property.

LEMMA 3.1. *If $\{(x^k, \tilde{\lambda}^k)\}$ is bounded, there exists a subsequence of $\{(x^k, \tilde{\lambda}^k)\}$ converging to some $(\bar{x}, \bar{\lambda})$ satisfying* (1.3) *if and only if $S^k \to +\infty$.*

Let α_i, $i = 1$ to 4, be some positive constants. The adaptation rule for p^k is as follows. An initial value p^1 is chosen; then

(3.7) Let l' be the index of the last iteration at which p^k has been changed ($l' = 1$ if this event never occurred). Then

$$p^k = \begin{cases} \tilde{\lambda}^k & \text{if } S^k > S^{l'} + \alpha_1, \\ p^{k-1} & \text{otherwise.} \end{cases}$$

We need some tools to define the adaptation law on r^k. For any $a > 0$, let

$$s(a) = \min\{10^q;\ a \leqq 10^q,\ q \text{ is an integer}\}.$$

If $\{a^n\}$ is a sequence of positive numbers, the transformed sequence $\{s(a^n)\}$ has the following properties:

$$\text{(3.8)} \quad \{\{a^n\} \to 0\} \Leftrightarrow \{s(a^n) \to 0\},$$

$$\text{(3.9)} \quad \{\limsup a^n = +\infty\} \Leftrightarrow (\limsup s(a^n) = +\infty\},$$

$$\text{(3.10)} \quad \{a^n \uparrow a,\ 0 < a < +\infty\} \Rightarrow \{s(a^n) = s(a) \text{ for } n \text{ large enough}\}.$$

We define the sequence ϕ^k by the following rule:

(0) $\phi^1 = 1$.

(1) Let l'' be the index of the last iteration at which $\phi^{l''}$ was nonnull.
If $S^k - S^{l''} > \alpha_2$ and $\rho^k \neq 1$, $\phi^k = S^k - S^{l''}$.
Else $\phi^k = 0$.

We obviously have

$$\lim S^k = +\infty \Rightarrow \left\{ \sum_k \phi^k = +\infty \text{ or } \rho^k = 1 \text{ for } k \text{ large enough} \right\}, \tag{3.11}$$

$$\lim S^k < +\infty \Rightarrow \phi^k = 0 \text{ for large enough.} \tag{3.12}$$

We suppose that

$$0 < \alpha_4 < 1. \tag{3.13}$$

The adaptation rule for r^k is given by

$$\begin{aligned} &\text{(i)} && r_1^k = (3+\alpha_3)\max\left(\|\tilde{\lambda}^k - p^k\|_D, \min\left(1, (D^k)^{\alpha_4}\right)\right), \\ &\text{(ii)} && r_2^k = \max\left(r_1^k, r_2^{k-1} - \phi^k\right), \\ &\text{(iii)} && r^k = s(r_2^k). \end{aligned} \tag{3.14}$$

The motivation for this rule is as follows. First we compute an estimate r_1^k that satisfies (2.4) if $\tilde{\lambda}^k$ is close to $\bar{\lambda}$; if the convergence occurs in the sense that $x^k \to \bar{x}$ and $D^k \to 0$, then we also have $r_1^k \geqq (D^k)^{\alpha_4}$; since $D^k / \|x^k - \bar{x}\|$ cannot converge to zero, hypothesis (2.5) is satisfied. Then in step (ii) we prevent r_2^k from decreasing if no sufficient progress has been made. Finally, in step (iii) we make a transformation in the hope to get r^k constant for k large enough when the convergence occurs (see case (b) of Theorem 3.1).

Remark 3.1. We may detail step (3) of Algorithm 1 as follows. k is set to $k+1$, then:

- First D^k, S^k, and ϕ^k are computed;
- Then p^k is computed using (3.7);
- Finally r^k is computed using (3.14), and H^k is set.

We now prove that the resulting algorithm is globally and superlinearly convergent.

THEOREM 3.1. *Let $\{x^k\}$ be a sequence computed by Algorithm* 1, *p^k and r^k being given by* (3.7) *and* (3.14). *We suppose that the sequences $\{H^k\}$ and $\{(H^k)^{-1}\}$ are bounded and that $0 < \sigma < \frac{1}{2}$. Then we have the following.*

(a) *One of the following three events occurs*:

(i) $\liminf_{k\to\infty} (\|\nabla f(x^k) + \nabla g(x^k)\tilde{\lambda}^k\|_U + \|g(x^k)\|) = 0$.

(ii) *For k large enough, (p^k, r^k) is equal to some (p, r), $\tilde{\lambda}^k$ is equal to λ^k and $\|\lambda^k - p\|_D < r$, $\theta_{p,r}(x^k) \to -\infty$ or $\{\nabla g(x^k)\}$ is unbounded.*

(iii) *For k large enough, p^k is equal to some p, $r^k \to +\infty$, $\limsup \|\tilde{\lambda}^k\|_D = +\infty$ and either $\{x^k\}$ is unbounded or some limit point $\bar{x}$ of $\{x^k\}$ does not satisfy* (1.2).

(b) *If $x^k \to \bar{x}$ satisfying* (1.2), *then there exists $\bar{\lambda}$ satisfying* (1.3) *and $(\tilde{\lambda}^k, d^k) \to (\bar{\lambda}, 0)$. If in addition $(\bar{x}, \bar{\lambda})$ satisfies* (2.1), *then hypotheses* (2.4) *and* (2.5) *are satisfied, and $\tilde{d}^k = d^k$ for k great enough. If* (1.7) *is also satisfied as well as* (2.1), *then r^k is equal to some $r > 0$ and $\rho^k = 1$ for k great enough, and x^k converges superlinearly to $\bar{x}$.*

Proof. (a) We suppose that (i) does not occur. Then S^k is bounded and ϕ^k is null for k large enough. Hence by (3.7) and (3.14), for k large enough, p^k is equal to some p and r^k is an increasing sequence.

If $\{\tilde{\lambda}^k\}$ is unbounded, (3.14)(i) implies that $r^k \to +\infty$. Then if $\{x^k\}$ is bounded, we deduce from (3.5) that it has a limit point $\bar{x}$ such that $\nabla g_i(\bar{x})$, $i = 1$ to m, are not linearly independent. Hence (iii) is satisfied.

If $\{\tilde{\lambda}^k\}$ is bounded, so is $\{r^k\}$ by (3.14); hence by (3.10), r^k is equal to some r for k large enough such that, by (3.14), $r > \|\tilde{\lambda}^k - p\|_D$. This implies that $\tilde{\lambda}^k = \lambda^k$ and $\tilde{d}^k = d^k$. Let us suppose that $\theta_{p,r}(x^k)$ is bounded from below and that $\{\nabla g(x^k)\}$ is bounded; we

must then get a contradiction to prove that (ii) holds. Let us prove that, for k large enough,

$$\theta_{p,r}(x^k) - \theta_{p,r}(x^k + \rho^k d^k) \geqq \sigma \rho^k (d^k)^t H^k d^k. \tag{3.15}$$

As $d^k = \tilde{d}^k$, we have, using the optimality conditions of (1.4):

$$\begin{aligned}
\theta_{p,r}(x^k) - \tilde{\theta}^k(x^k + d^k) &= -\nabla f(x^k)^t d^k - p^t \nabla g(x^k)^t d^k + r\|g(x^k)\|, \\
&= (d^k)^t H^k d^k + (\lambda^k - p)^t \nabla g(x^k)^t d^k + r\|g(x^k)\|, \\
&\geqq (d^k)^t H^k d^k + (r - \|\lambda^k - p\|_D)\|g(x^k)\|, \\
&\geqq (d^k)^t H^k d^k.
\end{aligned}$$

The convexity of $\tilde{\theta}$ implies

$$\theta_{p,r}(x^k) - \tilde{\theta}^k(x^k + \rho^k d^k) \geqq \rho^k(\tilde{\theta}^k(x^k) - \tilde{\theta}^k(x^k + d^k)) \geqq \rho^k (d^k)^t H^k d^k.$$

Then (3.15) is a consequence of the linesearch rule (3.3). Summing (3.15) over k, we deduce that

$$\sum_k \rho^k (d^k)^t H^k d^k < +\infty.$$

The equality $x^{k+1} - x^k = \rho^k d^k$ and the boundedness of $\{(H^k)^{-1}\}$ imply

$$\sum_k \|x^{k+1} - x^k\|_U \|d^k\|_U < +\infty. \tag{3.16}$$

If $d^k \to 0$ for some subsequence k, then the boundedness of $\{(H^k)\}$ and $\{\nabla g(x^k)\}$, relation (3.4)(ii), and the constraint of (1.4) imply that $S^k \to \infty$, which is impossible. If $\liminf \|d^k\| > 0$, then (3.16) implies the convergence of $\{x^k\}$ toward some $\bar{x}$. From the boundedness of $\{\lambda^k\}$, $\nabla g(x^k)$, and $\{(H^k)^{-1}\}$ and (3.4)(ii), we deduce that $\{d^k\}$ is bounded. As $\|x^{k+1} - x^k\| \to 0$ and $\liminf \|d^k\| > 0$, $\{\rho^k\}$ must converge to zero; this implies that for any integer $l \geqq 0$, the following inequality holds for k large enough:

$$\theta_{p,r}(x^k + (\beta)^l d^k) - \theta_{p,r}(x^k) > \sigma(\tilde{\theta}^k(x^k + (\beta)^l d^k) - \theta_{p,r}(x^k)).$$

Passing to the limit for some sequence K such that $\{d^k\}_{k \in K} \to \bar{d} \neq 0$, we get that, for all integers $l \geqq 0$,

$$\theta_{p,r}(\bar{x} + (\beta)^l \bar{d}) - \theta_{p,r}(\bar{x}) > \sigma(\tilde{\theta}(\bar{x} + (\beta)^l \bar{d}) - \theta_{p,r}(\bar{x})),$$

with $\tilde{\theta}$ defined as $\tilde{\theta}^k$ at point $\bar{x}$. Extracting from K a subsequence if necessary, we may suppose that for k in K, $\lambda^k \to \bar{\lambda}$ and $H^k \to \bar{H}$, with $\bar{H}$ positive definite. Passing to the limit in the constraints of (1.4) and (3.4)(ii), we find that $\bar{d}$ is the unique solution of the quadratic problem

$$\text{Minimize } \nabla f(\bar{x})^t d + \tfrac{1}{2} d^t \bar{H} d \quad \text{subject to } g(\bar{x}) + \nabla g(\bar{x})^t d = 0.$$

Hence the linesearch rule of Algorithm 1 starting at $\bar{x}$ must be satisfied at the first iteration by some $\bar{\rho} = (\beta)^{\bar{l}}$, in contradiction to the inequality above.

(b) If $x^k \to \bar{x}$ satisfying (1.2), then cases (ii) and (iii) of (a) may not occur; hence for some subsequences K, $D^k \to 0$ when $k \to \infty$ in K. As $\lim \nabla g(x^k) = \nabla g(\bar{x})$ has full rank, this implies that $\{\tilde{\lambda}^k\}_{k \in K}$ is bounded; as $\|\nabla f(x^k) + \nabla g(x^k)\tilde{\lambda}^k\|$ vanishes when $k \to \infty$ in K, any limit-point $\bar{\lambda}$ of $\{\tilde{\lambda}^k\}_{k \in K}$ satisfies (1.3); from (1.2) we obtain that $\bar{\lambda}$ is unique and $\{\tilde{\lambda}^k\}_{k \in K} \to \bar{\lambda}$.

By (3.7), p^k is changed only for values of D^k converging to zero, and hence for values of $\tilde{\lambda}^k$ converging to $\bar{\lambda}$; hence $p^k \to \bar{\lambda}$. But (see (3.5)) we obtain that, for all k,

$$\|\tilde{\lambda}^k - \bar{\lambda}\| \leqq \|\tilde{\lambda}^k - p^k\| + \|p^k - \bar{\lambda}\| \leqq \|\lambda^k - p^k\| + \|p^k - \bar{\lambda}\|.$$

Since $\{H^k\}$ and $\{(H^k)^{-1}\}$ are bounded, necessarily $d^k \to 0$ and $\lambda^k \to \bar{\lambda}$; with the above inequality, this implies that all the sequence $\tilde{\lambda}^k$ converges to $\bar{\lambda}$, and hence $D^k \to 0$. Now consider the case when (2.1) is also satisfied. Then the Jacobian of the optimality conditions (1.3) is nondegenerate at $(\bar{x}, \bar{\lambda})$; from it we deduce (see [2]) that there exists $C_1 > 0$ such that for k large enough

$$\|x^k - \bar{x}\| + \|\tilde{\lambda}^k - \bar{\lambda}\| \leqq C_1 D^k.$$

With (3.13) and (3.14), this implies that $r^k / \|x^k - \bar{x}\| \geqq r_1^k / \|x^k - \bar{x}\| > \min(1, (D^k)^{\alpha_4})/(C_1 D^k) \to +\infty$; hence (2.5) holds. Let us prove that (2.4) also holds with $\gamma = \alpha_3/2$; if not, there exists a subsequence K such that (using (3.14) for the left inequality), for k large enough in K,

$$(3+\alpha_3)\|\tilde{\lambda}^k - p^k\|_D \leqq r_1^k \leqq r^k \leqq (3+\alpha_3/2)\|\bar{\lambda} - p^k\|_D;$$

hence

$$(1+\alpha_3/3)\|\tilde{\lambda}^k - p^k\|_D \leqq (1+\alpha_3/6)\|\bar{\lambda} - p^k\|_D.$$

Using this inequality, we get

$$\begin{aligned}\|\bar{\lambda} - p^k\|_D &\leqq \|\bar{\lambda} - \tilde{\lambda}^k\|_D + \|\tilde{\lambda}^k - p^k\|_D, \\ &\leqq \|\bar{\lambda} - \tilde{\lambda}^k\|_D + \frac{1+\alpha_3/6}{1+\alpha_2/3}\|\bar{\lambda} - p^k\|_D.\end{aligned}$$

Hence for k large enough in K and for some $C_2 > 0$

$$\begin{aligned}\|\bar{\lambda} - p^k\|_D &\leqq \frac{1+\alpha_3/3}{\alpha_3/6}\|\bar{\lambda} - \tilde{\lambda}^k\|_D \\ &\leqq \frac{1\alpha_3/3}{\alpha_3/6} C_1 D^k \leqq C_2 (r^k)^{1/\alpha_4}.\end{aligned}$$

Hence for this subsequence $\|\bar{\lambda} - p^k\|_D / r^k$ converges to $+\infty$, contradicting our hypothesis; hence (2.4) holds.

If in addition (1.7) holds, let us prove that $\rho^k = 1$ for k large enough; this will prove that the superlinear convergence holds. If $\rho^k \neq 1$ for some subsequence K, from (3.11) and (3.14) we deduce that $r_1^k \to 0$ and $r_2^k \to 0$ (for all k). As $r^k \leqq 10 r_2^k$, $r^k \to 0$. As (2.4) and (2.5) are satisfied, r^k will be inferior to the value r_0 given by Theorem 2.1 with ε satisfying $\sigma = \frac{1}{2} - \varepsilon$. Then the conclusion of Theorem 2.1 implies that $\rho^k = 1$ for k large enough, in contradiction to our hypothesis. This proves that $\rho^k = 1$ and $\phi^k = 0$ for k large enough; hence by (3.14), as $r_1^k \to 0$, r_2^k and r^k are constant for k large enough. □

We comment on Theorem 3.1. Conclusion (a) is concerned with global convergence. Case (i) is the one in which the global convergence really occurs in a weak sense. Let us assume that $\nabla g(x)$ has full rank for all x. Then if (i) does not occur we deduce that $\{x^k\}$ is unbounded. An extension of the algorithm, including upper and lower bound constraints on all variables, might exclude this case (however, such an extension is not trivial). Conclusion (b) essentially says that when the convergence toward a regular local solution holds, the algorithm reduces to Newton's method (without linesearch) after a finite number of iterations. Our algorithm needs essentially

the computation of one quadratic program at each iteration (the amount of computation needed to update p^k and r^k is negligible). Also, the reduction of the nonsmooth part of the criterion might, even far from the optimum, improve the linesearch. The method is presently being tested on some large-scale network problems; numerical results will appear elsewhere.

Acknowledgments. The author thanks J. C. Dodu of Electricité de France and two referees for valuable suggestions that greatly improved the paper.

REFERENCES

[1] L. ARMIJO, *Minimization of function having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.

[2] P. T. BOGGS, J. W. TOLLE, AND P. WANG, *On the local convergence of quasi-Newton methods for constrained optimization*, SIAM J. Control Optim., 20 (1982), pp. 161-171.

[3] J. F. BONNANS, *Augmentability and exact penalizability in nonlinear programming under a weak second-order sufficiency condition*, Rept. 548, Inst. National de Recherche en Informatique et Automatique, Le Chesnay, France, 1986.

[4] J. F. BONNANS AND D. GABAY, *Une extension de la programmation quadratique successive*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Lecture Notes in Control and Inform. Sci. 63, Springer-Verlag, Berlin, New York, 1984, pp. 16-31.

[5] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multipliers Methods*, Academic Press, New York, 1982.

[6] R. M. CHAMBERLAIN, C. LEMARECHAL, H. C. PEDERSEN, AND M. J. D. POWELL, *The watchdog technique for forcing convergence in algorithm for constrained optimization*, Math. Programming Stud., 16 (1982), pp. 1-17.

[7] R. FLETCHER, *Practical Methods of Optimization*, Vol. 2, John Wiley, New York, Chichester, 1981.

[8] ———, *A model algorithm for composite non-differentiable optimization problems*, Math. Programming Stud., 17 (1982), pp. 67-76.

[9] D. GABAY, *Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization*, Math. Programming Stud., 16 (1982), pp. 18-44.

[10] S. P. HAN, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297-309.

[11] S. P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251-269.

[12] N. MARATOS, *Exact penalty function algorithms for finite dimensional and control optimization problems*, Ph.D. thesis, University of London, London, U.K., 1978.

[13] D. Q. MAYNE AND E. POLAK, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Stud., 16 (1982), pp. 45-61.

[14] M. J. D. POWELL, *A fast algorithm for nonlinearly constrained optimization calculations*, in Lecture Notes in Math. 630, G. A. Watson, ed., Springer-Verlag, Berlin, New York, 1978, pp. 144-157.

[15] ———, *The convergence of variable metric methods for nonlinearly constrained calculations*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 27-63.

[16] ———, *Methods for nonlinear constraints in optimization calculation*, Proc. IMA/SIAM Meeting on the State of the Art in Numerical Analysis, 1986.

[17] B. N. PSHENICHNYI AND Y. D. DANILIN, *Numerical Methods in Extremal Problems*, Mir, Moscow, 1977.

[18] K. TONE, *Revisions of constraint approximations in the successive Q.P. methods for nonlinear programming problems*, Math. Programming, 26 (1983), pp. 144-152.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 642-657, May 1989

011

CONTROL OF MARKOV CHAINS WITH LONG-RUN AVERAGE COST CRITERION: THE DYNAMIC PROGRAMMING EQUATIONS*

VIVEK S. BORKAR†

Abstract. The long-run average cost control problem for discrete time Markov chains on a countable state space is studied in a very general framework. Necessary and sufficient conditions for optimality in terms of the dynamic programming equations are given when an optimal stable stationary strategy is known to exist (e.g., for the situations studied in [*Stochastic Differential Systems, Stochastic Control Theory and Applications*, IMA Vol. Math. App. 10, Springer-Verlag, New York, Berlin, 1988, pp. 57-77]). A characterization of the desired solution of the dynamic programming equations is given in a special case. Also included is a novel convex analytic argument for deducing the existence of an optimal stable stationary strategy when that of a randomized one is known.

Key words. Markov chains, long-run average cost, optimal control, dynamic programming, stationary strategy

AMS(MOS) subject classifications. 60J10, 93E20

1. Introduction. In [5], the long-run average cost control problem for a Markov chain on a countable state space was studied under a very general setup. The two theoretical issues in this problem are: (i) establishing the existence of an optimal stable stationary strategy; and (ii) characterizing the same via the dynamic programming equations. The main thrust of [5] was (i), whereas (ii) was only cursorily touched on. The present paper has two objectives. One is to provide a more elegant alternative for a part of the argument leading to (i) in [5]. This alternative approach unmasks the underlying convex analytic structure not apparent in the lengthier argument of [5] and reflects the spirit of [6], where other cost criteria have been considered in a similar light. The principal objective of this paper, however, is to give a detailed treatment of the dynamic programming equations, settling (ii) above. The class of cost functions considered here is much more general than that of [5], where the cost functions have been assumed to be bounded.

Although this paper is a sequel to [5] in principle, it can be read independently.

The long-run average cost control problem for Markov chains dates back to [10] for the finite state space case and [7] for the countable state space. In most of its early development, the problem was treated as the "vanishing discount limit" of the discounted cost control problem. This classical approach is by now standard textbook material and the reader is referred to [2] and [13] (among others) for a succinct treatment. The shortcoming of this approach is that it needs a strong uniform stability condition in one of its various garbs [8]. This condition fails in many applications of interest such as controlled queues, as is evidenced in [12]. Motivated by this, in [3] and [4] Borkar has developed an alternative approach for Markov chains exhibiting a "nearest neighbor motion." The latter feature requires that each state have only finitely many neighbors and that the minimum path length from state i to any prescribed finite subset of the state space tend to infinity as i does. The approach was based on a characterization of the almost surely limit points of the empirical process of the joint state and control process. It was this approach that was carried over to a much more general setup in [5]. The present work complements [5] in the sense already described.

* Received by the editors June 8, 1987; accepted for publication (in revised form) October 3, 1988. This research was supported by National Science Foundation grant CDR-85-00108.

† Tata Institute of Fundamental Research, Bangalore Center, Bangalore 560012, India and Systems Research Center, University of Maryland, College Park, Maryland 20740.

The paper is organized as follows. Section 2 is devoted to a recapitulation of the notation introduced in [3] and used throughout [3]–[6]. This notation is not standard, but turns out to be extremely handy for the approach of [3]–[6] and the present paper. Using this notation, § 3 states the principal assumptions under which the dynamic programming equations will be studied and discusses various ramifications thereof. Section 4 treats the necessary conditions for optimality in terms of the dynamic programming equations. The proofs here are very much along the lines of those of [4, § 5] except for the extra work needed to take care of a possibly unbounded cost function and the absence of a "nearest neighbor motion" hypothesis. We include the full details to make this account self-contained. The so-called "value function" appearing in the dynamic programming equations is further studied in § 5. Section 6 establishes sufficient conditions for optimality using the dynamic programming equations. Section 7 concludes with a discussion of the problem of characterizing the desired solution of the dynamic programming equations. The Appendix describes the convex analytic argument mentioned at the beginning of this section.

Note that we develop the dynamic programming formalism given the existence of an optimal stable stationary strategy by independent means, e.g., those of [4] and [5]. This is the opposite of the conventional order of things.

2. Notation and preliminaries. Let X_n, $n = 1, 2, \cdots$, be a controlled Markov chain on state space $S = [1, 2, \cdots]$ with transition matrix $P_u = [[p(i, j, u_i)]]$, $i, j \in S$ indexed by the control vector $u = [u_1, u_2, \cdots]$. Here, $u_i \in D(i)$, $i \in S$, for some prescribed compact metric spaces $D(i)$. The functions $p(i, j, \cdot)$ are assumed to be continuous. By replacing each $D(i)$ by $\Pi D(k)$ and $p(i, j, \cdot)$ by its composition with the projection $\Pi D(k) \to D(i)$, we may assume that all $D(i)$'s are replicas of the same compact metric space D. We do so and then let L denote the countable product of copies of D with the product topology.

For any Polish space Y, $M(Y)$ will denote the space of probability measures on Y with the topology of weak convergence and for $n = 1, 2, \cdots, \infty$, Y^n will denote the n-times product of Y with itself.

A control strategy (CS) is a sequence $\{\xi_n\}$, $\xi_n = [\xi_n(1), \xi_n(2), \cdots]$ of L-valued random variables such that for $i \in S$, $n \geqq 1$,

$$P(X_{n+1} = i / X_m, \xi_m, m \leqq n) = p(X_n, i, \xi_n(X_n)). \tag{2.1}$$

We say that $\{X_n\}$ is governed by $\{\xi_n\}$ whenever (2.1) holds. If $\{\xi_n\}$ are identically distributed and ξ_n is independent of X_m, $m \leqq n$; ξ_m, $m < n$, for each n, we call the control strategy a stationary randomized strategy (SRS). We call it a stationary strategy (SS) if in addition to the above, the law of ξ_n, $n \geqq 1$, is assumed to be a Dirac measure. The motivation for this nomenclature is self-evident.

We assume throughout that S is a single communicating class under any SRS. If $\{X_n\}$ is positive recurrent under an SRS, we call the latter a stable SRS or SSRS. A stable SS (or SSS) is defined analogously.

Let $\{\xi_n\}$ be an SRS. Let $\Phi \in M(L)$ denote the common law of ξ_n, $n \geqq 1$. As we shall be interested only in the law of the process $(X_n, \xi_n(X_n))$, $n \geqq 1$, it suffices to consider Φ of the form $\Phi = \Pi \hat{\Phi}_i$, $\hat{\Phi}_i \in M(D)$ for $i \in S$. We shall denote this SRS by $\gamma[\Phi]$ and the corresponding transition matrix by $P[\Phi] = [[\int p(i, j, u) \hat{\Phi}_i(du)]]$. If the SRS is stable, it will have a unique invariant probability measure denoted by $\pi[\Phi] = [\pi[\Phi](1), \pi[\Phi](2), \cdots] \in M(S)$. For $f: S \to R$ and measurable $g: S \times D \to R$, define

$$C_f[\Phi] = \sum_i f(i) \pi[\Phi](i),$$

$$g_\Phi(i) = \int g(i, u)\hat{\Phi}_i(du), \qquad i \in S,$$

$$C_g[\Phi] = \sum_i g_\Phi(i)\pi[\Phi](i)$$

whenever the quantity on the right is defined. If $\Phi = \delta_\xi$ (i.e., the Dirac measure at ξ) for some $\xi \in L$, $\gamma[\Phi]$ is an SS and will be denoted by $\gamma\{\xi\}$. Correspondingly, we replace $P[\Phi]$, $\pi[\Phi]$, $C_f[\Phi]$, $C_g[\Phi]$ by $P\{\xi\} = P_\xi$, $\pi\{\xi\}$, $C_f\{\xi\}$, $C_g\{\xi\}$, respectively.

For an SSRS $\gamma[\Phi]$, define $\hat{\pi}[\Phi] \in M(S \times D)$ by $\int f d\hat{\pi}[\Phi] = C_f[\Phi]$ for all bounded continuous $f: S \times D \to R$. For an SSS $\gamma\{\xi\}$, define $\hat{\pi}\{\xi\} \in M(S \times D)$ analogously.

Let $k: S \times D \to R^+$ be continuous. Define

$$\psi_n = \frac{1}{n} \sum_{m=1}^{n} k(X_m, \xi_m(X_m)), \tag{2.2}$$

$$\psi_\infty = \liminf_{n \to \infty} \psi_n. \tag{2.3}$$

Our objective is to almost surely minimize ψ_∞ over all CS. If this is achieved for some CS, that CS will be said to be optimal.

Note that under an SSRS $\gamma[\Phi]$ or an SSS $\gamma\{\xi\}$, $\psi_n \to C_k[\Phi]$ almost surely ($\psi_n \to C_k\{\xi\}$ almost surely, respectively) where $+\infty$ is a possible value for $C_k[\Phi]$, $C_k\{\xi\}$. Our aim will be to show the existence of an optimal SSS and characterize the same. Thus it is natural to impose the condition that for at least one SSS $\gamma\{\xi\}$, $C_k\{\xi\} < \infty$. Let

$$\beta = \inf_{\text{SSRS}} C_k[\Phi], \qquad \alpha = \inf_{\text{SSS}} C_k\{\xi\}.$$

Then $\beta \leqq \alpha$.

Finally, let $\tau(i) = \min\{n > 1 | X_n = i\} (= \infty \text{ if } X_n \neq i, n \geqq 2)$, $i \in S$.

3. Stability under local perturbation. Consider the following two sets of assumptions:

(A1) $\liminf_{i \to \infty} \min_u k(i, u) \triangleq \eta > \beta$, (3.1)

(A2) $\sup_{\text{all CS}} E[\tau(1)^2 / X_1 = 1] < \infty$. (3.2)

Remarks. (a) It is not hard to see that in the absence of a blanket stability assumption, something like assumption (A1) would be needed to ensure the existence of an optimal SSS. Intuitively, (A) penalizes unstable behavior. For example, consider the case $k(i, u) = h(i)$ for some $h: S \to (0, \infty)$ satisfying $h(i) \to 0$ as $i \to \infty$. Then any SSS (or SSRS) yields a strictly positive cost while an unstable SS (or SRS) yields zero cost, making the latter optimal.

(b) More directly verifiable conditions that imply (3.2) are given in [5, § IX]. These either are conditions on the graph of the chain or require the existence of a suitable "Lyapunov function" (see [5] for details). An example appears in § 6.

In [5], it was proved that under Assumption (A1) or (A2) and for bounded k:

(1) $\psi_\infty \geqq \beta$ almost surely;

(2) There exists an SSRS $\gamma[\Phi]$ such that $C_k[\Phi] = \beta$;

(3) $\beta = \alpha$;

(4) There exists an SSS $\gamma\{\xi\}$ such that $C_k\{\xi\} = \alpha$.

In the Appendix, we provide an alternative argument to deduce (3) and (4) from (2).

In later sections we give necessary and sufficient conditions for an SSS $\gamma\{\xi\}$ to be optimal, using the "dynamic programming" equations. Some of these were stated without detailed proofs for bounded k in [5]. We make the following two assumptions:

(1) There exists an optimal SSS. (This would be implied, e.g., by Assumption (A1) or (A2).)

(2) (Stability under local perturbation.) If $\gamma\{\xi\}$ is an SSS satisfying $C_k\{\xi\}<\infty$, then for any $\xi' \in L$ such that $\xi'(i) \neq \xi(i)$ for at most one $i \in S$, $\gamma\{\xi'\}$ is an SSS and $C_k\{\xi'\}<\infty$.

In this section we make a few remarks about these conditions.

(i) If k is bounded, (2) is implied by the following condition. For any SSS $\gamma\{\xi\}$, $\gamma\{\xi'\}$ is an SSS whenever $\xi'(i)=\xi(i)$ for all but one $i \in S$.

(ii) If all SS are SSS and k is bounded, (2) holds trivially.

(iii) Condition (2) holds whenever each state in S has only finitely many neighbors, i.e., for each $i \in S$, there is a finite set $R_i \subset S$ such that $p(i, j, \cdot) \equiv 0$ for $j \notin R_i$. To see this, pick $i=1$, for example. If $\gamma\{\xi\}$ is an SSS and $C_k\{\xi\}<\infty$,

$$\infty > E[\tau(1)/X_1=1] \geqq P(\tau(j)<\tau(1)/X_1=1)E[\tau(1)/X_1=j]$$

for each $j \neq 1$. Since $P(\tau(j)<\tau(1)/X_1=1)>0$ for $j \neq 1$ by positive recurrence and single communicating class hypothesis, $a_j = E[\tau(1)/X_1=j]<\infty$ for all $j \neq 1$, and hence for all j. Thus under $\gamma\{\xi'\}$,

$$E[\tau(1)/X_1=1]=1+\sum_{j \in R_1 \setminus \{1\}} P(1, j, \xi'(1))a_j < \infty.$$

A similar argument using the fact that

$$E\left[\sum_{m=1}^{\tau(1)-1} k(X_m, \xi(X_m))/X_1=1\right]<\infty$$

under $\gamma\{\xi\}$ shows that under $\gamma\{\xi'\}$,

$$E\left[\sum_{m=1}^{\tau(1)-1} k(X_m, \xi'(X_m))/X_1=1\right]<\infty.$$

(iv) The following example describes a situation where (2) fails. Relabel S as $\{a_{00}, a_{10}, a_{11}, a_{20}, a_{21}, a_{22}, a_{30}, a_{31}, a_{32}, a_{33}, a_{40}, \cdots\}$. Let $D=[1.5, 3]$. Let $p(i, j, u)=1$, for all $u \in D$, $i=a_{mn}$, $j=a_{m(n+1)}$, $m=1, 2, \cdots$, $n=0, 1, \cdots, m-1$, and for $i=a_{mm}$, $j=a_{00}$, $m=1, 2, \cdots$. Let $f(\alpha)=\sum_n n^{-\alpha}$ for $\alpha \in D$ and $p(a_{00}, a_{m0}, \alpha)=f(\alpha)^{-1}m^{-\alpha}$, $m=1, 2, \cdots$.

Let $\{X_n\}$ be a Markov chain governed by the SS that chooses the control α whenever the chain is in a_{00}. (The transition probabilities for all transitions except those out of a_{00} are control-independent.) Letting $\tau=\inf\{n>1 | X_n=a_{00}\}$, we have

$$E[\tau/X_1=a_{00}]=f(\alpha)^{-1}\sum_{m=1}^{\infty}(m+2)m^{-\alpha},$$

which is finite for $\alpha \in (2, 3]$ and ∞ for $\alpha \in [1.5, 2]$. Changing D to $[2.5, 3]$, we get an example where (2) holds, but a_{00} has infinitely many neighbors. Thus the condition in (iii) above is sufficient but not necessary for (2) to be true.

(v) In what follows, we shall often construct a new SS from a given SSS by changing finitely many of its components. By (2), it will be an SSS and by an assumption already made, S will be a single communicating class under it.

As an immediate consequence of these assumptions, we have the following lemma.

LEMMA 3.1. *Let $\gamma\{\xi\}$ be an* SSS *for which $C_k\{\xi\}<\infty$. Then for any $i \in S$, $u \in D$,*

$$\sum_{j \in S} p(i, j, u)E_\xi\left[\sum_{n=1}^{\tau(1)} k(X_m, \xi(X_m))/X_1=j\right]<\infty, \tag{3.3}$$

$$\sum_{j\in S} p(i,j,u)E_\xi[\tau(1)/X_1=j]<\infty, \tag{3.4}$$

where $E_\xi[\ ]$ *denotes the expectation under* $\gamma\{\xi\}$.

Proof. Note that

$$\infty > E_\xi\left[\sum_{n=1}^{\tau(1)} k(X_n,\xi(X_n))/X_1=1\right] \geqq aE_\xi\left[\sum_{n=1}^{\tau(1)} k(X_n,\xi(X_n))/X_1=j\right]$$

where

$$a = P(\{X_n, n\geqq 2\} \text{ hits } j \text{ before hitting } 1/X_1=1)>0.$$

Thus

$$E_\xi\left[\sum_{n=1}^{\tau(1)} k(X_n,\xi(X_n))/X_1=j\right]<\infty \quad \forall j\in S. \tag{3.5}$$

Similarly,

$$E_\xi[\tau(1)/X_1=j]<\infty \quad \forall j\in S. \tag{3.6}$$

Let $\{\xi'_n\}$ denote a CS such that $\xi'_n=\xi$ for $n\geqq 2$ and $\xi'_1(i)=u$ for some fixed $i\in S$, $u\in D$. Let $\{X'_n\}\{X_n\}$ be the chains governed by $\{\xi'_n\}$, $\gamma\{\xi\}$, respectively, with $X'_1=X_1=i$. Let $\tau'(i)=\inf\{n>1|X'_n=i\}$. Then

$$\begin{aligned}
&k(i,u)+\sum_{j\in S} p(i,j,u)E_\xi\left[\sum_{n=1}^{\tau(1)} k(X_n,\xi(X_n))/X_1=j\right]\\
&\quad=E\left[\sum_{n=1}^{\tau'(1)} k(X'_n,\xi'_n(X'_n))\right]\\
&\quad=E\left[\left(\sum_{n=1}^{\tau'(1)} k(X'_n,\xi'_n(X'_n))\right)I\{\tau'(1)<\tau'(i)\}\right]\\
&\qquad+E\left[\left(\sum_{n=1}^{\tau'(1)} k(X'_n,\xi'_n(X'_n))\right)I\{\tau'(1)>\tau'(i)\}\right].
\end{aligned}$$

When we define $\varphi\in L$ by $\varphi(j)=\xi(j)$ for $i\neq j$ and $\varphi(i)=u$, the above is

$$\begin{aligned}
&\leqq E_\varphi\left[\sum_{n=1}^{\tau(1)} k(X_n,\varphi(X_n))/X_1=i\right]+E_\varphi\left[\sum_{n=1}^{\tau(i)} k(X_n,\varphi(X_n))/X_1=i\right]\\
&\qquad+E_\xi\left[\sum_{n=1}^{\tau(1)} k(X_n,\xi(X_n))/X_1=i\right]<\infty
\end{aligned}$$

by virtue of (3.5). Condition (3.3) follows. Condition (3.4) follows from (3.6) by analogous arguments. □

4. Necessary conditions for optimality. This section proves necessary conditions for the optimality of an SSS in terms of the dynamic programming equations (Theorem 4.1 below). Let $\gamma\{\xi\}$ be an SSS with $C_k\{\xi\}<\infty$. Define $V\{\xi\}=[V\{\xi\}(1), V\{\xi\}(2),\cdots]^T$ by

$$V\{\xi\}(i)=E_\xi\left[\sum_{n=1}^{\tau(1)-1}(k(X_n,\xi(X_n))-C_k\{\xi\})/X_1=i\right], \qquad i\in S.$$

This is well defined by virtue of (3.5), (3.6). By Lemma 3.1,

$$\sum_{j\in S} p(i,j,u)V\{\xi\}(j)$$

is also well defined for $u \in D$. Let $1_c = [1, 1, \cdots]^T$, $U =$ the infinite identity matrix $[[\delta_{ij}]]$ and $Q_\xi = [k(1, \xi(1)), k(2, \xi(2)), \cdots]^T$. The following lemma is recalled from [4].

LEMMA 4.1. *$V\{\xi\}(1) = 0$ and*

$$C_k\{\xi\}1_c = (P\{\xi\} - U)V\{\xi\} + Q_\xi. \tag{4.1}$$

Proof. The first claim follows from the fact that

$$C_k\{\xi\} = \sum_{i \in S} \pi\{\xi\}(i)k(i, \xi(i)) = E_\xi\left[\sum_{n=1}^{\tau(1)-1} k(X_n, \xi(X_n))/X_1 = 1\right] \Big/ E[\tau(1) - 1/X_1 = 1].$$

Since $V\{\xi\}(1) = 0$, we have

$$V\{\xi\}(i) = k(i, \xi(i)) - C_k\{\xi\} + E_\xi\left[\left(\sum_{n=2}^{\tau(1)-1} (k(X_n, \xi(X_n)) - C_k\{\xi\})\right) I\{\tau(1) > 2\}/X_1 = i\right]$$

$$= k(i, \xi(i)) - C_k\{\xi\} + E[V\{\xi\}(X_2) I\{\tau(1) > 2\}/X_1 = i]$$

$$= k(i, \xi(i)) - C_k\{\xi\} + E[V\{\xi\}(X_2)/X_1 = i]$$

$$= k(i, \xi(i)) - C_k\{\xi\} + \sum_{j \in S} p(i, j, \xi(i)) V\{\xi\}(j)$$

for $i \in S$. Equation (4.1) follows. □

Let $A \subset S$ be a finite set containing 1 and $\xi' \in L$ such that $\xi'(i) = \xi(i)$ for $i \notin A$. Let A_n, $n = 1, 2, \cdots$ be an increasing family of finite subsets of S containing A and increasing to S. Define $\sigma_m = \min\{n \geqq 1 | X_n \notin A_m\}$, $m = 1, 2, \cdots$ and $\sigma = \min\{n \geqq 1 | X_n \in A\}$.

Observe that by the assumptions of the previous section, $\gamma\{\xi'\}$ is an SSS with $C_k\{\xi'\} < \infty$.

LEMMA 4.2. $\lim_{n \to \infty} E_{\xi'}[V\{\xi\}(X_{\sigma_n}) I\{\tau(1) \geqq \sigma_n\}/X_1 = 1] = 0$.

Proof. For $i \notin A$,

$$V\{\xi\}(i) = E_\xi\left[\sum_{n=1}^{\sigma-1} (k(X_n, \xi(X_n)) - C_k\{\xi\})/X_1 = i\right] + E_\xi\left[\sum_{n=\sigma}^{\tau(1)-1} (k(X_n, \xi(X_n)) - C_k\{\xi\})/X_1 = i\right],$$

where we use the fact that $V\{\xi\}(1) = 0$. The first term on the right remains unchanged if $E_\xi[\ ]$ is replaced by $E_{\xi'}[\ ]$ and $k(X_n, \xi(X_n))$ by $k(X_n, \xi'(X_n))$. The second term is bounded in absolute value by

$$K = \max_{i \in A} E_\xi\left[\sum_{n=1}^{\tau(1)-1} (k(X_n, \xi(X_n)) + C_k\{\xi\})/X_1 = i\right].$$

Let $c = \max(C_k\{\xi\}, C_k\{\xi'\})$. Then for $i \notin A$,

$$|V\{\xi\}(i)| \leqq E_{\xi'}\left[\sum_{n=1}^{\sigma-1} (k(X_n, \xi'(X_n)) + c)/X_1 = i\right] + K$$

$$\leqq E_{\xi'}\left[\sum_{n=1}^{\tau(1)-1} (k(X_n, \xi'(X_n)) + c)/X_1 = i\right] + K.$$

Hence

$$|E_{\xi'}[V\{\xi\}(X_{\sigma_n})I\{\tau(1)\geqq\sigma_n\}/X_1=1]|$$
$$\leqq E_{\xi'}\left[\left(\sum_{m=\sigma_n}^{\tau(1)-1}(k(X_m,\xi'(X_m))+c)\right)I\{\tau(1)\geqq\sigma_n\}/X_1=1\right]$$
$$+KE_{\xi}[I\{\tau(1)\geqq\sigma_n\}/X_1=1]\to 0 \quad \text{as } n\to\infty. \qquad \square$$

THEOREM 4.1. *If $\gamma\{\xi\}$ is an optimal* SSS, *then*

$$\beta 1_c=\min_{\varphi}((P\{\varphi\}-U)V\{\xi\}+Q_{\varphi}). \tag{4.2}$$

Remark. By (4.1), it follows that the minimum in (4.2) is attained at $\varphi=\xi$.

Proof. Suppose not. Note that the ith component of $(P\{\varphi\}-U)V\{\xi\}+Q_{\varphi}$ depends only on the ith component of φ. Thus there exist $i\in S$, $u\in D$ and $\Delta>0$ such that if $\varphi\in L$ is defined by $\varphi(j)=\xi(j)$ for $j\neq i$ and $\varphi(i)=\mu$, then

$$\beta 1_c=(P\{\varphi\}-U)V\{\xi\}+Q_{\varphi}+J, \tag{4.3}$$

$J=[0,0,\cdots,0,\Delta,0,\cdots,0]$, with Δ in the ith place. Let $\{X_n\}$ be the chain governed by $\gamma\{\varphi\}$ with $X_1=i$. We may take $i=1$ by relabeling S if necessary. By our assumptions of the preceding section, $\gamma\{\varphi\}$ is an SSS with $C_k\{\varphi\}<\infty$. Set $A=\{1\}$ and $\{A_n\}$ as above. By (4.3),

$$\beta=E_{\varphi}[V\{\xi\}(X_{m+1})/X_m]-V\{\xi\}(X_m)+k(X_m,\varphi(X_m))+\Delta I\{X_m=1\}.$$

Thus for $n\geqq 1$,

$$\begin{aligned}\beta(\tau(1)\wedge\sigma_n-1)=&\sum_{m=1}^{\tau(1)\wedge\sigma_n-1}(E_{\varphi}[V\{\xi\}(X_{m+1})/X_m]-V\{\xi\}(X_m))\\&+\sum_{m=1}^{\tau(1)\wedge\sigma_n-1}k(X_m,\varphi(X_m))+\Delta\sum_{m=1}^{\tau(1)\wedge\sigma_n-1}I\{X_m=1\}.\end{aligned} \tag{4.4}$$

Since $V\{\xi\}(X_{\tau(1)})=V\{\xi\}(1)=0$, we have

$$E\left[\sum_{m=1}^{\tau(1)\wedge\sigma_n-1}(E_{\varphi}[V\{\xi\}(X_{m+1})/X_m]-V\{\xi\}(X_m))\right]$$
$$=-E\left[\sum_{m=1}^{\tau(1)\wedge\sigma_n}(V\{\xi\}(X_m)-E_{\varphi}[V\{\xi\}(X_m)/X_{m-1}])\right]$$
$$+E[V\{\xi\}(X_{\sigma_n})I\{\tau(1)\geqq\sigma_n\}]$$
$$=E[V\{\xi\}(X_{\sigma_n})I\{\tau(1)\geqq\sigma_n\}] \quad \text{(by the optional sampling theorem)}$$
$$\to 0 \quad \text{as } n\to\infty,$$

by Lemma 4.2. Taking expectations in (4.4), letting $n\to\infty$, and then dividing by $E_{\varphi}[\tau(1)]-1$, we get

$$\beta=C_k\{\varphi\}+\Delta\pi\{\varphi\}(1)>C_k\{\varphi\},$$

contradicting the definition of β. The claim follows. $\square$

The function $i\to V\{\xi\}(i)$ corresponding to an optimal SSS $\gamma\{\xi\}$ is called the value function and (4.2) the dynamic programming equations.

5. The value function. The definition of the value function in the preceding section depends on our choice of the specific optimal SSS $\gamma\{\xi\}$ and the state "1." In this section, we eliminate this dependence.

LEMMA 5.1. *Suppose* $W=[W(1), W(2), \cdots]^T$ *satisfies*

$$\beta 1_c = \inf_{\varphi}((P\{\varphi\}-U)W+Q_{\varphi}),$$

$$\sup_i |W(i)-V\{\xi\}(i)|<\infty$$

for some optimal SSS $\gamma\{\xi\}$. *Then* $W=V\{\xi\}+$ *constant* $\times 1_c$. *In particular, if* $W(1)=0$, *then* $W=V\{\xi\}$.

Proof. Since

$$\beta 1_c=(P\{\xi\}-U)V\{\xi\}+Q_{\xi}\leqq(P\{\xi\}-U)W+Q_{\xi},$$

we have

$$(P\{\xi\}-U)(W-V\{\xi\})\geqq 0.$$

It follows that under $\gamma\{\xi\}$, $W(X_n)-V\{\xi\}(X_n)$, $n\geqq 1$, is a bounded submartingale with respect to the natural filtration of $\{X_n\}$ and hence converges. Since $\{X_n\}$ visits each $i\in S$ infinitely often, this is possible only if $W(X_n)-V\{\xi\}(X_n)$, $n\geqq 1$, is almost surely a constant sequence. □

LEMMA 5.2. *Let* $\gamma\{\xi\}$, *let* $V\{\xi\}$ *be as above and for some* $i\in S$, *and define* $V'\{\xi\}=[V'\{\xi\}(1), V'\{\xi\}(2), \cdots]^T$ *by*

$$V'\{\xi\}(j)=E_{\xi}\left[\sum_{m=1}^{\tau(i)-1}(k(X_m,\xi(X_m))-\beta)/X_1=j\right].$$

Then

$$V'\{\xi\}=V\{\xi\}+\text{constant}\times 1_c.$$

Remark. Note that (4.2) remains unchanged if we change $V\{\xi\}$ by a constant multiple of 1_c.

Proof. For any $j\in S$,

$$|V\{\xi\}(j)-V'\{\xi\}(j)|\leqq E\left[\sum_{m=\tau(1)\wedge\tau(i)}^{\tau(1)\vee\tau(i)}(k(X_m,\xi(X_m))+\beta)/X_1=j\right]$$

$$\leqq E\left[\sum_{m=1}^{\tau(1)}(k(X_m,\xi(X_m))+\beta)/X_1=i\right] \tag{5.1}$$

$$+E\left[\sum_{m=1}^{\tau(i)}(k(X_m,\xi(X_m))+\beta)/X_1=1\right].$$

Since the choice of state "1" in the preceding section was arbitrary, it is clear that (4.2) also holds with $V'\{\xi\}$ replacing $V\{\xi\}$. The claim now follows from (5.1) and the preceding lemma. □

LEMMA 5.3. *For* $i\in S$, $\gamma\{\xi\}$, $V\{\xi\}$ *as above,*

$$V\{\xi\}(i)=\min E_{\varphi}\left[\sum_{m=1}^{\tau(1)-1}(k(X_m,\varphi(X_m))-\beta)/X_1=i\right],$$

where the minimum is over all SSS $\gamma\{\varphi\}$.

Proof. For $i=1$, the claim follows from the fact that for any SSS $\gamma\{\varphi\}$,

$$E_\varphi\left[\sum_{m=1}^{\tau(1)-1}(k(X_m,\varphi(X_m))-\beta)/X_1=1\right]=E_\varphi[\tau(1)-1](C_k\{\varphi\}-\beta)\geqq 0=V\{\xi\}(1).$$

Take $i\neq 1$. Suppose the claim is false. Then for some SSS $\gamma\{\varphi\}$,

$$E_\varphi\left[\sum_{m=1}^{\tau(1)-1}(k(X_m,\varphi(X_m))-\beta)/X_1=i\right]<V\{\xi\}(i). \tag{5.2}$$

Consider the chain $\{X_n\}$ with $X_1=1$ governed by a CS $\{\xi_n\}$ such that between each successive returns to state 1, $\xi_n=\xi$ until $\{X_n\}$ hits i (if it does) and equals φ from then on until it returns to 1. From (5.2) it follows that under $\{\xi_n\}$,

$$\begin{aligned}
&E\left[\sum_{m=1}^{\tau(1)-1}(k(X_m,\xi_m(X_m))-\beta)\right]\\
&\quad=E\left[\left(\sum_{m=1}^{\tau(1)-1}(k(X_m,\xi(X_m))-\beta)\right)I\{\tau(i)>\tau(1)\}\right]\\
&\qquad+E\left[\left(\sum_{m=1}^{\tau(i)-1}(k(X_m,\xi(X_m))-\beta)\right)I\{\tau(i)<\tau(1)\}\right]\\
&\qquad+E\left[\sum_{m=\tau(i)}^{\tau(1)-1}(k(X_m,\phi(X_m))-\beta)I\{\tau(i)<\tau(1)\}\right]<\text{first two terms}\\
&\qquad+E_\xi\left[\left(\sum_{m=\tau(i)}^{\tau(1)-1}(k(X_m,\xi(X_m))-\beta)\right)I\{\tau(i)<\tau(1)\}\right]\\
&\quad=V\{\xi\}(1)=0,
\end{aligned}$$

where the strict inequality holds because, under the positive recurrence and single communicating class conditions, $P_\xi(\tau(i)<\tau(1)/X_1=1)>0$.

Letting $\{\tau_n\}$ denote the successive return times to 1, we can easily see that

$$\sum_{m=\tau_i}^{\tau_{i+1}-1}(k(X_m,\xi_m(X_m))-\beta),\qquad i\geqq 1,$$

are independently and identically distributed. Thus by the strong law of large numbers,

$$\begin{aligned}
&\left[\sum_{m=1}^{\tau_n}(k(X_m,\xi_m(X_m))-\beta)\right]\Big/\tau_n\\
&\quad\xrightarrow{\text{a.s.}}E\left[\sum_{m=1}^{\tau(1)-1}(k(X_m,\xi_m(X_m))-\beta)/X_1=1\right]\Big/E[\tau(1)-1]<0.
\end{aligned}$$

This contradicts the optimality of $\gamma\{\xi\}$, proving the claim. □

COROLLARY 5.1. *$V\{\xi\}$ above does not depend on our choice of a specific optimal* SSS $\gamma\{\xi\}$.

6. Sufficient conditions for optimality. In this section, we develop sufficient conditions for optimality in terms of the dynamic programming equations. Let $\gamma\{\xi_0\}$ be an optimal SSS and let $V\{\xi_0\}$ be the value function. The traditional form of the sufficient conditions is as follows.

THEOREM 6.1. *Suppose an* SSS $\gamma\{\xi\}$ *satisfies* $C_k\{\xi\}<\infty$ *and*

$$\beta 1_c=(P\{\xi\}-U)V\{\xi_0\}+Q_\xi. \tag{6.1}$$

Then $\gamma\{\xi\}$ is optimal.

Proof. We argue as in the proof of Theorem 4.1 to conclude that

$$\beta E_\xi[\tau(1)\wedge\sigma_n - 1/X_1 = 1] = E_\xi\left[\sum_{m=1}^{\tau(1)\wedge\sigma_n - 1} k(X_m, \xi(X_m))/X_1 = 1\right] + E_\xi[V\{\xi_0\}(X_{\sigma_n})I\{\tau(1) > \sigma_n\}/X_1 = 1].$$

By Lemma 5.3, the last term is dominated by

$$E_\xi\left[E_\xi\left[\sum_{m=0}^{\tau(1)-\sigma_n-1}(k(X_{\sigma_n+m}, \xi(X_{\sigma_n+m})) - \beta)\right]\Big/ X_{\sigma_n}\right] I\{\tau(1) > \sigma_n\}/X_1 = 1\Bigg]$$
$$= E_\xi\left[\left(\sum_{m=0}^{\tau(1)-\sigma_n-1}(k(X_{\sigma_n+m}, \xi(X_{\sigma_n+m})) - \beta)\right) I\{\tau(1) > \sigma_n\}/X_1 = 1\right]$$
$$\to 0 \quad \text{as } n \to \infty.$$

Thus

$$\beta \geqq E_\xi\left[\sum_{m=1}^{\tau(1)-1} k(X_m, \xi(X_m))/X_1 = 1\right]\Big/ E_\xi[\tau(1) - 1/X_1 = 1] = C_k\{\xi\}.$$

Since $\beta \leqq C_k\{\xi\}$ in any case, $\beta = C_k\{\xi\}$ and the claim follows. □

We shall consider another variant of this. Call an SSS $\gamma\{\xi\}$ locally optimal if $C_k\{\xi\} \leqq C_k\{\xi'\}$ for all ξ' for which $\xi(i) \neq \xi'(i)$ for at most finitely many i.

THEOREM 6.2. *Suppose an* SSS $\gamma\{\xi\}$ *satisfies*

$$C_k\{\xi\} < \infty, \qquad C_k\{\xi\}1_c = \min_\varphi ((P\{\varphi\} - U)V\{\xi\} + Q_\varphi). \tag{6.2}$$

Then $\gamma\{\xi\}$ *is locally optimal.*

Proof. Let $\xi' \in L$ be such that $\xi'(i) \neq \xi(i)$ for at most finitely many i. Then $\gamma\{\xi'\}$ is an SSS by our hypothesis (2) of § 3. Let $\{X_n\}$ be a chain governed by $\gamma\{\xi'\}$ with $X_1 = 1$. By (6.2),

$$C_k\{\xi\} \leqq E_{\xi'}[V\{\xi\}(X_{n+1})/X_n] - V\{\xi\}(X_n) + k(X_n, \xi'(X_n))$$

for $n \geqq 1$. As in the proof of Theorem 4.1, we can prove that for $n \geqq 1$,

$$C_k\{\xi\}E_{\xi'}[\tau(1)\wedge\sigma_n - 1] \leqq E_{\xi'}\left[\sum_{m=1}^{\tau(1)\wedge\sigma_n - 1} k(X_m, \xi'_m(X_m))\right] + E_{\xi'}[V\{\xi\}(X_{\sigma_n})I\{\tau(1) \geqq \sigma_n\}].$$

The last term on the right tends to zero as $n \to \infty$ by Lemma 4.2. Thus letting $n \to \infty$ in the above and then dividing through by $E_{\xi'}[\tau(1) - 1]$, we get $C_k\{\xi\} \leqq C_k\{\xi'\}$. □

Remark. The converse can also be proved along the lines of Theorem 4.1.

COROLLARY 6.1. *Suppose all locally optimal* SSS *are optimal. Then an* SSS $\gamma\{\xi\}$ *is optimal if and only if* (6.2) *holds.*

COROLLARY 6.2. *Suppose all* SS *are* SSS *and* $\{\pi\{\xi\} | \xi \in L\}$ *is tight in* $M(S)$. *In addition, suppose that k is bounded. Then an* SSS $\gamma\{\xi\}$ *is optimal if and only if* (6.2) *holds.*

Proof. Let $\gamma\{\xi_0\}$ be an optimal SSS and $\gamma\{\xi\}$ a locally optimal SSS. Define $\xi^n \in L$ by $\xi^n(i) = \xi_0(i)$ for $i \leqq n$, $= \xi(i)$ for $i > n$. Then $P\{\xi^n\} \to P\{\xi_0\}$ termwise. Let $\pi\{\xi^n\} \to \pi \in M(S)$ along a subsequence. By Scheffe's Theorem, this convergence is also in total variation. Thus letting $n \to \infty$ along this subsequence in the equation $\pi\{\xi^n\}P\{\xi^n\} = \pi\{\xi^n\}$, we get $\pi P\{\xi_0\} = \pi$, i.e., $\pi = \pi\{\xi_o\}$. Thus $\hat{\pi}\{\xi^n\} \to \hat{\pi}\{\xi_0\}$ and hence

$$C_k\{\xi^n\} \to C_k\{\xi_o\} \leqq C_k\{\xi\}.$$

But $C_k\{\xi\} \leqq C_k\{\xi^n\}$ by local optimality. Thus $C_k\{\xi\} = C_k\{\xi_o\}$ and $\gamma\{\xi\}$ is optimal. The claim follows from the preceding corollary. □

Clearly, any example that satisfies the hypotheses of Corollary 6.2 will also be an example for Corollary 6.1. One example where the hypotheses of the former are satisfied

is a stable chain where each state has finitely many neighbors as in (iii) of § 4, with the control affecting the transitions out of at most finitely many states. A more serious example can be constructed involving chains that show a strong uniform "drift" toward a finite set when they are sufficiently away from it; the set must be uniform with respect to all the controls. We do this next.

Let $\infty > K > \varepsilon > 0$ be given. Let a_n, $n \geqq 1$, be a sequence in $[2\varepsilon, K]$. Define $W: S \to R_+$ by $W(1) = a_1$, $W(n) = a_1 + \cdots + a_n$, $n \geqq 2$. Consider a controlled Markov chain $\{X_n\}$ whose transition matrix P_u satisfies the following:

(i) $p(i, j, \cdot) \equiv 0$ for $j \neq i+1$ or $i-1$,

> 0 otherwise for $i \in S \backslash \{1\}$,

$p(1, j, \cdot) \equiv 0$ for $j > 2$, > 0 otherwise.

(ii) $p(i, i+1, \cdot) \leqq (a_i - \varepsilon)/(a_{i+1} + a_i)$, $i \geqq 1$.

Letting $\mathcal{F}_n = \sigma(X_m, \xi_m, m \leqq n)$, it is then easy to check that

$$E[(W(X_{n+1}) - W(X_n)) I\{X_n \geqq 2\}/\mathcal{F}_n] < -\varepsilon I\{X_n \geqq 2\}.$$

As observed in [5, pp. 72–73], it then follows from the results of [9, § 2] that the hypotheses of Corollary 6.2 hold.

The proof of Corollary 6.2 indicates that the requirements can be relaxed to the following. For any two SSS $\gamma\{\xi'\}$ and $\gamma\{\xi''\}$, the set $\{\pi\{\xi\} | n \geqq 1$ such that $\xi(i) = \xi'(i)$ for $i \leqq n$ and equals $\xi''(i)$ for $i > n\}$ is tight. This would be true, e.g., if the map $\xi \to \pi\{\xi\}$ from the set $\{\xi \in L | \gamma\{\xi\}$ is an SSS$\}$ to $M(S)$ were continuous. This seems an eminently reasonable thing to expect in "typical" situations. Thus it is tempting to conjecture that local optimality implies optimality, except possibly in some highly pathological situations. Unfortunately, at the moment we are unable to capture the essence of exactly what is involved.

Note that whenever local optimality implies optimality, (6.2) is a much better condition for optimality than (6.1), because all the quantities involved depend only on the SSS $\gamma\{\xi\}$ under scrutiny and no prior knowledge of β or $V\{\xi_0\}$ is needed.

7. Characterizing the solution of the dynamic programming equations. By a solution of the dynamic programming equations, we mean a pair (c, W), $c \in R^+$, $W = [W(1), W(2), \cdots]^T$ an infinite column vector, such that

$$c1_c = \min_{\xi} ((P\{\xi\} - U) W + Q_\xi). \tag{7.1}$$

Clearly, $(\beta, V\{\xi_0\})$ in the foregoing is one solution. Note that if (c, W) is a solution, so is $(c, W + \Delta 1_c)$ for any $\Delta \in R$.

In this section, we shall give a characterization that isolates the distinguished solution $(\beta, V\{\xi_0\})$ from among the solution set for the special case when (A1) holds.

LEMMA 7.1. *Under* (A1), $V\{\xi_0\}(i)$, $i \in S$, *is bounded from below.*

Proof. By (A1), $A = \{i \in S | k(i, u) < \beta$ for some $u \in D(i)\}$ is a finite set. Let $\sigma = \min\{n \geqq 1 | X_n \in A\}$. Then for $i \in S$,

$$V\{\xi_0\}(i) = E_{\xi_0}\left[\sum_{m=1}^{\tau(1)-1} (k(X_m, \xi_0(X_m)) - \beta)/X_1 = i\right]$$

$$\geqq E_{\xi_0}\left[\sum_{m=\sigma}^{\tau(1)-1} (k(X_m, \xi_0(X_m)) - \beta) I\{\tau(1) > \sigma\}/X_1 = i\right]$$

$$\geqq -\beta \sum_{j \in A} E_{\xi_0}[\tau(1)/X_1 = j]. \qquad \square$$

Let $G=\{f\colon S\to R \mid f(1)=0 \text{ and } \inf_i f(i)>-\infty\}$.

LEMMA 7.2. *Let* $(c, W)\in R^+\times G$ *be a solution of* (7.1). *Then* $c\geqq\beta$.

Proof. Let $\varepsilon>0$. Then there exists an SS $\gamma\{\xi\}$ such that

$$(\varepsilon+c)1_c\geqq(P\{\xi\}-U)W+Q_\xi. \tag{7.2}$$

Let $\{X_n\}$ be a chain governed by $\gamma\{\xi\}$ with $X_1=1$. Summing up $X_1, X_2, \cdots, X_n$ rows of (7.2), we have

$$\begin{aligned}(c+\varepsilon)n &\geqq E_\xi[W(X_{n+1})/X_n]\\ &\quad-\sum_{m=2}^{n}(W(X_m)-E_\xi[W(X_m)/X_{m-1}])+\sum_{m=1}^{n}k(X_m,\xi(X_m)).\end{aligned} \tag{7.3}$$

By familiar arguments, we obtain

$$\begin{aligned}(c+\varepsilon)n &\geqq E_\xi[W(X_{n+1})/X_1]+E_\xi\left[\sum_{m=1}^{n}k(X_m,\xi(X_m))/X_1\right]\\ &\geqq K+E_\xi\left[\sum_{m=1}^{n}k(X_m,\xi(X_m))/X_1\right],\end{aligned}$$

where $K>-\infty$ is a lower bound on $W(i)$, $i\in S$. Divide by n and let $n\to\infty$. If $\gamma\{\xi\}$ is not an SSS, then it is not hard to deduce from (A1) that

$$\liminf_{n\to\infty}\frac{1}{n}E_\xi\left[\sum_{m=1}^{n}k(X_m,\xi(X_m))/X_1=1\right]\geqq\eta.$$

Thus

$$c+\varepsilon\geqq\eta\geqq\beta.$$

If $\gamma\{\xi\}$ is an SSS, then

$$\liminf_{n\to\infty}\frac{1}{n}E_\xi\left[\sum_{m=1}^{n}k(X_m,\xi(X_m))/X_1=1\right]\geqq C_k\{\xi\}\geqq\beta.$$

Since $\varepsilon>0$ is arbitrary, the claim follows. □

LEMMA 7.3. *If* (β, W), $W\in G$, *is a solution of* (7.1), *then* $W\geqq V\{\xi_0\}$ *termwise.*

Proof. Let $0<\varepsilon<\eta-\beta$. Then there exists an SS $\gamma\{\xi\}$ such that

$$(\beta+\varepsilon)1_c\geqq(P\{\xi\}-U)W+Q_\xi. \tag{7.4}$$

If $\gamma\{\xi\}$ is not an SSS, we may argue as in the proof of the preceding lemma to conclude that $\beta+\varepsilon\geqq\eta$, a contradiction. Hence $\gamma\{\xi\}$ is an SSS. Let $\{\xi_n\}$ be a chain governed by $\gamma\{\xi\}$ with $X_1=i$ for some $i\in S$. For $\{\sigma_n\}$ as before, we can deduce from (7.4) that

$$\begin{aligned}W(i)\geqq&\sum_{m=1}^{\tau(1)\wedge\sigma_n-1}(k(X_m,\xi(X_m))-(\beta+\varepsilon))\\ &-\sum_{m=2}^{\tau(1)\wedge\sigma_n}(W(X_m)-E_\xi[W(X_m)/X_{m-1}])+W(X_{\tau(1)\wedge\sigma_n}).\end{aligned} \tag{7.5}$$

By Fatou's Lemma,

$$\begin{aligned}\liminf_{n\to\infty}E_\xi[W(X_{\tau(1)\wedge\sigma_n})/X_1=i]&\geqq\liminf_{n\to\infty}E_\xi[W(X_{\sigma_n})I\{\tau(1)>\sigma_n\}/X_1=i]\\ &\geqq 0.\end{aligned}$$

Thus taking expectations in (7.5) and letting $n \to \infty$,

$$W(i) \geqq E_\xi\left[\sum_{m=1}^{\tau(1)-1} (k(X_m, \xi(X_m)) - (\beta+\varepsilon))/X_1 = i\right].$$

Since ε can be made arbitrarily close to zero,

$$\begin{aligned} W(i) &\geqq E_\xi\left[\sum_{m=1}^{\tau(1)-1} (k(X_m, \xi(X_m)) - \beta)/X_1 = i\right], \\ &\geqq V\{\xi_0\}(i) \end{aligned}$$

where the last inequality follows from Lemma 5.3. □

LEMMA 7.4. *For W as above, $W = V\{\xi_0\}$.*

Proof. Let $K > 0$ be a finite number such that $W(i) \geqq -K$ for $i \in S$. Then for each $i \in S$,

$$\sum_{j \in S} p(i, j, u) W(j) + k(i, u) = \sum_{j \in S} p(i, j, u)(W(j) + K) + k(i, u) - K.$$

The first term on the right is a monotone increasing limit of continuous functions in the variable u, and hence is lower semicontinuous in u. Since $k(i, \cdot)$ is continuous, the left-hand side above is lower semicontinuous in u, and hence attains a minimum at some $u_i \in D$. Let $\xi = [u_1, u_2, \cdots] \in L$. Then

$$\beta 1_c = (P\{\xi\} - U) W + Q_\xi.$$

By the arguments used in the proof of Lemma 7.2, $\gamma\{\xi\}$ is an SSS. Since

$$\beta 1_c \leqq (P\{\xi\} - U) V\{\xi_0\} + Q_\xi,$$

we have

$$(P\{\xi\} - U)(W - V\{\xi_0\}) \leqq 0.$$

Letting $\{X_n\}$ be a chain governed by $\gamma\{\xi\}$ with $X_1 = 1$, this and Lemma 7.3 imply that $V = W - V\{\xi_0\}$ satisfies

$$V(X_n) \geqq o = V(X_1), \qquad E_\xi[V(X_{n+1})/X_n] \leqq V(X_n),$$

$n = 1, 2, \cdots$. This is possible only if $V(X_n) = o$ almost surely for each n. Since $\gamma\{\xi\}$ is an SSS, $X_n = i$ infinitely often almost surely for each $i \in S$. Hence $V(i) = 0$ for $i \in S$. The claim follows. □

The following theorem summarizes the above.

THEOREM 7.1. *Among all solutions (c, W) of (7.1) in $R^+ \times G$, $(\beta, V\{\xi_o\})$, is the unique solution corresponding to the least value of c.*

Appendix A. Recall (1)-(4) in the beginning of § 3. Here, we derive (3) and (4) from (2) using a simple convex analytic argument.

LEMMA A.1. *The set $B = \{\hat{\pi}[\Phi] | \gamma[\Phi]$ an SSRS$\}$ is convex and closed in $M(S \times D)$.*

Proof. Let $\gamma[\Phi_1]$, $\gamma[\Phi_2]$ be two SSRS with $\Phi_1 = \Pi_i \hat{\Phi}_{1i}$, $\Phi_2 = \Pi_i \hat{\Phi}_{2i}$. Let $0 \leqq a \leqq 1$ and define $\Phi = \Pi_i \hat{\Phi}_i$ by

$$\hat{\Phi}_i = (a\pi[\Phi_1](i)\hat{\Phi}_{1i} + (1-a)\pi[\Phi_2](i)\hat{\Phi}_{2i})/(a\pi[\Phi_1](i) + (1-a)\pi[\Phi_2](i)), \qquad i \in S.$$

From this definition and the fact that

$$\pi[\Phi_i] P[\Phi_i] = \pi[\Phi_i], \qquad i = 1, 2,$$

it is easily seen that

$$(a\pi[\Phi_1] + (1-a)\pi[\Phi_2]) P[\Phi] = (a\pi[\Phi_1] + (1-a)\pi[\Phi_2]).$$

Thus

$$\pi[\Phi] = a\pi[\Phi_1] + (1-a)\pi[\Phi_2].$$

Hence

$$\pi[\Phi](i)\hat{\Phi}_i = a\pi[\Phi_1](i)\hat{\Phi}_{1i} + (1-a)\pi[\Phi_2](i)\hat{\Phi}_{2i},\ i \in S,$$

implying

$$\hat{\pi}[\Phi] = a\hat{\pi}[\Phi_1] + (1-a)\hat{\pi}[\Phi_2].$$

The convexity follows. Now let $\gamma[\Phi_n]$, $n = 1, 2, \cdots$, be SSRS such that $\hat{\pi}[\Phi_n] \to \hat{\pi}$ for some $\hat{\pi} \in M(S \times D)$. Let $\pi \in M(S)$, $\pi = [\pi(1), \pi(2), \cdots]$, be the image of $\hat{\pi}$ under the projection $S \times D \to S$. Then $\pi[\Phi_n] \to \pi$ in $M(S)$. Disintegrate $\hat{\pi}$ as $\hat{\pi}(\{i\} \times A) = \pi(i)\varphi_i(A)$, $i \in S$, A a Borel subset of D, where $\varphi_i \in M(D)$ for each i. Define $\varphi \in L$ by $\varphi = \Pi_i \varphi_i$. Since $p(\cdot, j, \cdot)$, $j \in S$, are continuous,

$$\int p(\cdot, j, \cdot)\, d\hat{\pi}[\Phi_n] \to \int p(\cdot, j, \cdot)\, d\hat{\pi}, \qquad j \in S.$$

Thus

$$\pi[\Phi_n]P[\Phi_n] \to \pi P[\varphi]$$

termwise. Since $\pi[\Phi_n] \to \pi$ and $\pi[\Phi_n]P[\Phi_n] = \pi[\Phi_n]$, for $n \geqq 1$, we have $\pi P[\varphi] = \pi$, i.e., $\pi = \pi[\varphi]$. Hence $\hat{\pi} = \hat{\pi}[\varphi]$ and we are done. □

Let $\gamma[\Phi]$, $\Phi = \Pi_i \hat{\Phi}_i$ be an SSRS such that for some $i_o \in S$ and $0 < a < 1$, there exist φ_1, φ_2 in $M(D)$ such that

$$\text{(A1)} \quad \begin{aligned} &\int p(i_0, \cdot, u)\hat{\Phi}_{i_0}(du) = a \int p(i_0, \cdot, u)\varphi_1(du) + (1-a) \int p(i_0, \cdot, u)\varphi_2(du), \\ &\int p(i_0, \cdot, u)\varphi_1(du) \neq \int p(i_0, \cdot, u)\varphi_2(du) \end{aligned}$$

as vectors, the integrations being termwise. Without loss of generality, we shall assume that $i_0 = 1$.

LEMMA A.2. *$\hat{\pi}[\Phi]$ is not an extreme point of B.*

Proof. Define $\Phi_i \in M(L)$ by $\Phi_i = \varphi_i \times \Pi_{j=2}^{\infty} \hat{\Phi}_j$, $i = 1, 2$. Let τ, τ_1, τ_2 denote the first return time to 1 under $\gamma[\Phi]$, $\gamma[\Phi_1]$, $\gamma[\Phi_2]$, respectively, when the chain starts at 1. It is easily seen that

$$\begin{aligned} E[\tau] &= 1 + \sum_{j \neq 1} \int p(1, j, u)\hat{\Phi}_1(du)E[\tau / X_1 = j] \\ &= a\left(1 + \sum_{j \neq 1} \int p(1, j, u)\varphi_1(du)E[\tau / X_1 = j]\right) \\ &\qquad + (1-a)\left(1 + \sum_{j \neq 1} \int p(1, j, u)\varphi_2(du)E[\tau / X_1 = j]\right) \\ &= aE[\tau_1] + (1-a)E[\tau_2]. \end{aligned}$$

Since $E[\tau] < \infty$, $E[\tau_i] < \infty$ for $i = 1, 2$, implying that $\gamma[\Phi_i]$, $i = 1, 2$, are SSRS. If $\pi[\Phi] = \pi[\Phi_1] = \pi[\Phi_2]$, the equation

$$\sum_k \pi[\Phi_i](k) \int p(k, j, u)\hat{\Phi}_{ik}(du) = \pi[\Phi_i](j), \qquad i = 1, 2,$$

contradicts (A1) for some j. Hence any two of $\pi[\Phi]$, $\pi[\Phi_1]$, $\pi[\Phi_2]$ are distinct from each other. Let $b \in (0, 1)$ be such that

$$a = b\pi[\Phi_1](1)/(b\pi[\Phi_1](1) + (1-b)\pi[\Phi_2](1)).$$

We argue as in the proof of the preceding lemma to conclude that

$$\hat{\pi}[\Phi] = b\hat{\pi}[\Phi_1] + (1-b)\hat{\pi}[\Phi_2].$$

The claim follows. □

COROLLARY A.1. *The extreme points of B are of the form $\hat{\pi}\{\xi\}$, where $\xi \in L$ satisfies the following*:

(*) *For each $i \in S$, $p(i, \cdot, \xi(i))$ is an extreme point of $\{p(i, \cdot, u) | u \in D\} \subset M(S)$.*

THEOREM A.1. *If an optimal* SSRS *exists, an optimal* SSS *satisfying* (*) *exists. (In particular, $\beta = \alpha$.)*

Proof. Let $\gamma[\Phi]$ be an optimal SSRS. Let $\bar{S} = SU\{\infty\}$ be the one-point compactification of S. We may view B as a subset of $M(\bar{S} \times D)$ by identifying each element of $M(S \times D)$ with that element of $M(\bar{S} \times D)$ that coincides with it when restricted to $S \times D$ and has zero mass at $\{\infty\} \times D$. Let $\bar{B}$ denote the closure of B in $M(\bar{S} \times D)$. Viewing $\hat{\pi}[\Phi]$ as an element of $\bar{B}$, Choquet's Theorem [11] implies that $\hat{\pi}[\Phi]$ is the barycenter of a probability measure ν supported on the set of extreme points of $\bar{B}$. Since $\hat{\pi}[\Phi]$ has no mass at $\{\infty\} \times D$, ν is almost surely supported on the set of extreme points of B itself. Letting E denote the latter set, we have

$$\int_E \left(\int k \, d\hat{\pi} \right) \nu(d\hat{\pi}) = C_k[\Phi].$$

Thus if there is no $\hat{\pi} \in E$ such that $\int k \, d\hat{\pi} = C_k[\Phi]$, there would necessarily exist a $\hat{\pi} \in E$ for which $\int k \, d\hat{\pi} < C_k[\Phi]$. By the preceding corollary, each $\hat{\pi} \in E$ is of the form $\hat{\pi}\{\xi\}$ for some SSS $\gamma\{\xi\}$ satisfying (*). Thus we have a contradiction to the optimality of $\gamma[\Phi]$ unless $C_k[\Phi] = C_k\{\xi\}$ for some SSS $\gamma\{\xi\}$ with $\hat{\pi}\{\xi\} \in E$. □

Remark. As in [5], we can prove that (A1) or (A2) imply (1) and (2) above. (Although k is assumed to be bounded in [5], this part of the arguments of [5] goes through without any difficulty for the more general k's considered here.) The above can then replace the arguments of [5] to deduce (3) and (4) from (1) and (2). This alternative approach is both simpler and says more.

REFERENCES

[1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] D. P. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[3] V. S. BORKAR, *Controlled Markov chains and stochastic networks*, SIAM J. Control Optim., 21 (1983), pp. 652-666.

[4] ———, *On minimum cost per unit time control of Markov chains*, SIAM J. Control Optim., 22 (1984), pp. 965-978.

[5] ———, *Control of Markov chains with long-run average cost criterion*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. Fleming and P. L. Lions, eds., IMA Vol. Math. Appl. 10, Springer-Verlag, Berlin, New York, 1988, pp. 57-77.

[6] ———, *A convex analytic approach to Markov decision processes*, Probab. Theory Related Fields, 78 (1988), pp. 583-602.

[7] C. DERMAN, *Denumerable state Markovian decision processes—average cost criterion*, Ann. of Math. Statist., 37 (1966), pp. 1545-1554.

[8] A. FEDERGRUEN, A. HORDIJK, AND H. C. TIJMS, *A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices*, J. Appl. Probab., 15 (1978), pp. 842-847.

[9] B. HAJEK, *Hitting-time and occupation-time bounds implied by drift analysis with applications*, Adv. Appl. Probab., 14 (1982), pp. 502–525.
[10] R. HOWARD, *Dynamic Programming and Markov Decision Processes*, MIT Press, Cambridge, MA, 1960.
[11] R. PHELPS, *Lectures on Choquet's Theorem*, Van Nostrand, New York, 1966.
[12] Z. ROSBERG, P. VARAIYA, AND J. WALRAND, *Optimal control of service in tandem queues*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 600–609.
[13] S. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 3, pp. 658-672, May 1989

012

RANK INVARIANTS OF NONLINEAR SYSTEMS*

M. D. DI BENEDETTO†, J. W. GRIZZLE‡, AND C. H. MOOG§

Abstract. A linear algebraic framework for the analysis of rank properties of nonlinear systems is introduced. This framework gives a high-level interpretation of several existing algorithms built around the recursive computation of certain algebraic ranks associated with right-invertibility, left-invertibility, and dynamic decoupling. Furthermore, it can be used to establish links between these algorithms and the differential algebraic approach, as well as to solve some static and dynamic noninteracting control problems.

Key words. invertibility, decoupling, zeros at infinity, differential algebra, nonlinear systems analysis

AMS(MOS) subject classifications. primary 93C10; secondary 93B25

1. Introduction.

Consider a nonlinear control system of the following form:

$$\Sigma: \quad \dot{x} = f(x) + g(x)u, \tag{1.1a}$$

$$y = h(x) \tag{1.1b}$$

where, for simplicity, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$, and $f(\cdot)$, the columns of $g(\cdot)$ and the rows of $h(\cdot)$ are *meromorphic* functions of x (on all of $\mathbb{R}^n$); that is, they are elements of the fraction field $\mathcal{F}(x)$ of the ring of analytic functions of x. Our goal is to associate to such a system a chain of (nondifferential) vector spaces and show that it contains a rich amount of structural information about the system. More precisely, the subspaces will recover in a unified way, the inversion algorithm of Singh [1], the generic ranks of Nijmeijer [2], the dynamic decoupling algorithms of Descusse and Moog [3] and Nijmeijer and Respondek [4], and the differential output rank of Fliess [5], [6]. The approach adopted in the paper has been largely inspired by the differential vector spaces considered in [7] by Fliess.

To proceed, suppose that the input function $u(t)$ to the system (1.1) is N times continuously differentiable. Then by Taylor's Theorem,

$$u(t) = \sum_{k=0}^{N} u^{(k)} \frac{(t-t_0)^k}{k!} + R_N(t-t_0),$$

where t_0 is some initial point in time, $u^0 := u(t_0)$, $u^{(i+1)} := d/dt\, u^{(i)}(t)|_{t=t_0}$, and R_N is a remainder term. View $x, u, \cdots, u^{(n-1)}$ as variables and let $\mathcal{K}$ denote the field consisting of the set of rational functions of $(u, \cdots, u^{(n-1)})$ with coefficients that are meromorphic in x. Recall that given such a field, say in the variables $\nu = (\nu_1, \cdots, \nu_j)$, we define $\partial/\partial\nu_i$ acting on a meromorphic function $\eta(\nu) = \pi(\nu)/\theta(\nu)$, where $\pi(\cdot)$ and $\theta(\cdot)$ are analytic, by the usual quotient rule of calculus,

$$\frac{\partial}{\partial \nu_i}\frac{\pi(\nu)}{\theta(\nu)} := \frac{\left(\theta(\nu)\dfrac{\partial}{\partial \nu_i}\pi(\nu) - \pi(\nu)\dfrac{\partial}{\partial \nu_i}\theta(\nu)\right)}{\theta^2(\nu)}. \tag{1.2}$$

* Received by the editors April 25, 1988; accepted for publication (in revised form) October 3, 1988.

† Instituto Universitario Navale, Istituto di Matematica, Via Acton, 38, 80133 Naples, Italy and Dipartimento di Informatica e Sistemistica, Universítà di Roma, "La Sapienza," Rome, Italy.

‡ Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122. The work of this author was supported by the National Science Foundation under contract NSF ECS-88-96136.

§ Laboratoire d'Automatique de Nantes, Ecole Nationale Superieure de Mecanique, Unité Associée au Centre National de la Recherche Scientifique, 1, rue de la Noë, 44072 Nantes Cedex 03, France. The work of this author was performed while he was a NATO visiting professor at the University of Michigan.

Then we define the differential of η by

$$d\eta(\nu) := \sum_{i=1}^{j} \frac{\partial \eta(\nu)}{\partial \nu_i} d\nu_i. \tag{1.3}$$

Returning to the system (1.1), we define in a natural way,

$$\dot{y} = \dot{y}(x, u) = \frac{\partial h}{\partial x}[f(x) + g(x)u], \tag{1.4a}$$

$$\begin{aligned} y^{(k+1)} &= y^{(k+1)}(x, u, \cdots, u^{(k)}) \\ &= \frac{\partial y^{(k)}}{\partial x}[f(x) + g(x)u] + \sum_{i=0}^{k-1} \frac{\partial y^{(k)}}{\partial u^{(i)}} u^{(i+1)}. \end{aligned} \tag{1.4b}$$

Note that $\dot{y}, \cdots, y^{(n)}$ so defined have components in the field $\mathcal{K}$.

Let $\mathcal{E}$ denote the vector space (over $\mathcal{K}$) spanned by $\{dx, du, \cdots, du^{(n-1)}\}$. It is essential to remark that this is an ordinary, nondifferential vector space as opposed to the setting proposed in [7]. Now we introduce the chain of subspaces $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$ of $\mathcal{E}$ by

$$\mathcal{E}_k := \operatorname{span}\{dx, d\dot{y}, \cdots, dy^{(k)}\} \tag{1.5}$$

and define the associated list of dimensions $\rho_0 \leqq \cdots \leqq \rho_n$ by

$$\rho_k := \dim \mathcal{E}_k. \tag{1.6}$$

It is important to note that in (1.5) and (1.6) the span and dimension are both taken with respect to the field $\mathcal{K}$, and *not* the real numbers. Hence ρ_k is a well-defined integer and is not a function of $x, u, \cdots, u^{(n-1)}$. Note also that we abuse notation slightly because $\mathcal{E}_0 := \operatorname{span}\{dx, dy\}$, which is easily seen to be equal to $\operatorname{span}\{dx\}$ since the output y only depends on x. Finally, in the above, as well as in all that follows, "d" of a vector or vector valued function means "d" of each of its components; that is, $\mathcal{E}_0 = \operatorname{span}\{dx_1, \cdots, dx_n\}$, etc.

In the sequel we argue that the chain of subspaces $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$ gives a linear algebraic framework that clarifies many structural properties of nonlinear systems and leads to a synthesis of many previous works on rank invariants of nonlinear systems [1]–[9].

2. Four concepts becomes one. For a linear system, there are many equivalent approaches to defining (or characterizing) its rank. More or less, all possible approaches have been extended to nonlinear systems in an effort to understand such properties as right-invertibility, left-invertibility, dynamic decoupling, etc. In general, these extensions lead to distinct notions when broadened to the class of nonlinear systems [1], [10]. However, we will show that the linear algebraic framework of § 1 can be used to establish the equivalence of four of them. A first attempt at this, using cruder tools, was made in [11].

2.1. Jacobian matrices. In [2], Nijmeijer considers systems of the form (1.1) where $f(\cdot)$, the columns of $g(\cdot)$ and the rows of $h(\cdot)$ are *analytic* functions of x. He defines $\dot{y}, \cdots, y^{(n)}$ as before, and introduces

$$J_k(x, u, \cdots, u^{(k-1)}) := \frac{\partial(\dot{y}, \cdots, y^{(k)})}{\partial(u, \cdots, u^{(k-1)})}. \tag{2.1}$$

J_k is an analytic function of its arguments; hence, we can define

$$R_k := \text{generic rank (over the real numbers) of } J_k(x, u, \cdots, u^{(k-1)}). \tag{2.2}$$

Nijmeijer shows that $R_{k+1} - R_k$ is a nondecreasing sequence of integers and says that (1.1) is right-invertible if $R_n - R_{n-1} = p$, the number of scalar output components. He goes on to relate the integers R_k, for a restricted class of systems, to a set of integer invariants coming from the Δ^*-algorithm; that set, in analogy with results from linear system theory, was termed the "structure at infinity."

We now relate the list of generic ranks $\{R_k\}$ to the list of integers $\{\rho_k\}$ defined in (1.6). The first step is to observe that R_k is equal to the dimension of the row span of J_k, where the field is taken as $\mathcal{K}$. That is, if we define

$$\mathcal{V}_k := \operatorname{span}\left\{\frac{\partial \dot{y}}{\partial u}\, du, \cdots, \sum_{i=0}^{k-1} \frac{\partial y^{(k)}}{\partial u^{(i)}}\, du^{(i)}\right\}, \tag{2.3}$$

then $R_k = \dim \mathcal{V}_k$. It follows immediately that for $k \geqq 1$

$$\mathcal{E}_k = \operatorname{span}\{dx\} \oplus \mathcal{V}_k; \tag{2.4}$$

hence, we have the following theorem.

THEOREM 2.1. $\rho_k = n + R_k$, $1 \leqq k \leqq n$.

This yields the following corollary.

COROLLARY 2.2. *System* (1.1) *is right invertible in the sense of Nijmeijer if and only if* $\dim \mathcal{E}_n - \dim \mathcal{E}_{n-1} = p$, *the number of scalar output components.*

In closing this subsection, we note that the subspaces $\mathcal{V}_k$ give a linear algebraic interpretation of the Jacobian matrices J_k. In the case considered in this section where f, g, and h are analytic, both $\mathcal{E}_k$ and $\mathcal{V}_k$ can be viewed as analytic codistributions on the manifold $M \times T^{(n-1)}U$, where $T^{(n-1)}U$ is the $(n-1)$th order tangent bundle [12, Chap. 1] of the input manifold $U = \mathbb{R}^m$. However, it is easy to see that $\mathcal{E}_k$ is always involutive, whereas $\mathcal{V}_k$, the projection of $\mathcal{E}_k$ along span $\{dx\}$, does not in general enjoy this property. This gives us some reason to believe that the $\mathcal{E}_k$'s are more intrinsic, and certainly indicates that they are more amenable to analysis.

2.2. The inversion algorithm. In [1], Singh introduces an algorithm for calculating the left-inverse of a nonlinear system; his algorithm is in fact a generalization of previous algorithms, due to Silverman [13] and Hirschorn [14], that are only applicable under some restrictive conditions. This algorithm has since been taken up by different authors and has been used to define a finite zero structure for nonlinear systems [15] (important for certain stabilization problems), and a structure at infinity [9] (important for noninteracting control problems and model matching).

In the following, it will be shown that the inversion algorithm actually constructs, step by step, a *basis* for the chain of subspaces $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$. This shows that the chain $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$ contains all of the above-cited structural information yielded by the inversion algorithm, and also confirms the earlier claim that the chain $\mathcal{E}_0 \subset \cdots \subset \mathcal{E}_n$ embodies important structural information on a high level, independently of any particular algorithmic considerations.

The inversion algorithm as detailed in [15] is now given, with the exception that, instead of allowing a large class of analytic transformations, we will use a particular meromorphic transformation. This idea was first sketched in [9].

Step 1. Calculate

$$\dot{y} = \frac{\partial h}{\partial x}[f(x) + g(x)u] \tag{2.5}$$

and write it as

$$\dot{y} =: a_1(x) + b_1(x)u. \tag{2.6}$$

Define

$$s_1 := \operatorname{rank} b_1(x), \tag{2.7}$$

where the rank is taken over the field of meromorphic functions of x. Permute, if necessary, the components of the output so that the first s_1 rows of $b_1(x)$ are linearly independent. Decompose y as

$$\dot{y} = \begin{pmatrix} \dot{\tilde{y}}_1 \\ \dot{\hat{y}}_1 \end{pmatrix}, \tag{2.8}$$

where $\dot{\tilde{y}}_1$ consists of the first s_1 rows of $\dot{y}$. Since the last rows of $b_1(x)$ are linearly dependent upon the first s_1 rows, we can write

$$\dot{\tilde{y}}_1 = \tilde{a}_1(x) + \tilde{b}_1(x)u, \tag{2.9a}$$

$$\dot{\hat{y}}_1 = \hat{y}_1(x, \dot{\tilde{y}}_1), \tag{2.9b}$$

where the last equation is affine in $\dot{\tilde{y}}_1$. Finally, set $\tilde{B}_1(x) := \tilde{b}_1(x)$.

Step $k+1$. Suppose that in Steps 1 through k, $\dot{\tilde{y}}_1, \cdots, \tilde{y}_k^{(k)}, \hat{y}_k^{(k)}$ have been defined so that

$$\begin{aligned} &\dot{\tilde{y}}_1 = \tilde{a}_1(x) + \tilde{b}_1(x)u, \\ &\vdots \\ &\tilde{y}_k^{(k)} = \tilde{a}_k(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k-1,\ i \leqq j \leqq k\}) \\ &\qquad + \tilde{b}_k(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k-1,\ i \leqq j \leqq k-1\})u, \\ &\hat{y}_k^{(k)} = \hat{y}_k^{(k)}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k\}) \end{aligned} \tag{2.10}$$

and so that they are rational functions of $\tilde{y}_i^{(j)}$ with coefficients in the field of meromorphic functions of x. Suppose also that the matrix $\tilde{B}_k := [\tilde{b}_1^T, \cdots, \tilde{b}_k^T]^T$ has full rank equal to s_k. Then calculate

$$\hat{y}_k^{(k+1)} = \frac{\partial}{\partial x}\hat{y}_k^{(k)}[f(x) + g(x)u] + \sum_{i=1}^{k}\sum_{j=i}^{k} \frac{\partial \hat{y}_k^{(k)}}{\partial \tilde{y}_i^{(j)}} \tilde{y}_i^{(j+1)} \tag{2.11}$$

and write it as

$$\hat{y}_k^{(k+1)} = a_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k+1\}) + b_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k\})u. \tag{2.12}$$

Define $B_{k+1} := [\tilde{B}_k^T, b_{k+1}^T]^T$, and

$$s_{k+1} := \operatorname{rank} B_{k+1}, \tag{2.13}$$

where the rank is taken with respect to the field of rational functions of $\{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k-1,\ i \leqq j \leqq k\}$ with coefficients in the field of meromorphic functions of x. Permute, if necessary, the components of $\hat{y}_k^{(k+1)}$ so that the first s_{k+1} rows of B_{k+1} are linearly independent. Decompose $\hat{y}_k^{(k+1)}$ as

$$\hat{y}_k^{(k+1)} = \begin{pmatrix} \tilde{y}_{k+1}^{(k+1)} \\ \hat{y}_{k+1}^{(k+1)} \end{pmatrix}, \tag{2.14}$$

where $\tilde{y}_{k+1}^{(k+1)}$ consists of the first $(s_{k+1} - s_k)$ rows. Since the last rows of $B_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k\})$ are linearly dependent on the first s_{k+1} rows, we can write

$$\begin{aligned} &\dot{\tilde{y}}_1 = \tilde{a}_1(x) + \tilde{b}_1(x)u, \\ &\vdots \\ &\tilde{y}_{k+1}^{(k+1)} = \tilde{a}_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k+1\}) \\ &\qquad + \tilde{b}_{k+1}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k,\ i \leqq j \leqq k\})u, \\ &\hat{y}_{k+1}^{(k+1)} = \hat{y}_{k+1}^{(k+1)}(x, \{\tilde{y}_i^{(j)} \mid 1 \leqq i \leqq k+1,\ i \leqq j \leqq k+1\}), \end{aligned} \tag{2.15}$$

where once again everything is rational in $\tilde{y}_i^{(j)}$. Finally, set $\tilde{B}_{k+1} := [\tilde{B}_k^T, \tilde{b}_{k+1}^T]^T$.

End of Step $k+1$.

It is now possible to state and prove the main result of this subsection.

THEOREM 2.3. *For each* $1 \leqq k \leqq n$:

(a) $\{dx, \{d\tilde{y}_i^{(j)} | 1 \leqq i \leqq k, i \leqq j \leqq k\}\}$ *is a basis for* $\mathscr{E}_k$;

(b) $\dim \mathscr{E}_k = n + s_1 + \cdots + s_k$.

Proof. Part (b) is an immediate consequence of (a), which will be proved by induction. For $k = 1$, the statement is obvious. Suppose that (a) holds at Step k; it will now be shown that it also holds at Step $k+1$. By construction, $\tilde{B}_{k+1}$ is a rational function of $\{\tilde{y}_i^{(j)} | 1 \leqq i \leqq k, i \leqq j \leqq k\}$ with coefficients in the field of meromorphic functions of x. As $\tilde{y}_i^{(j)}$ is a rational function of $(u, \cdots, u^{(n-1)})$ with coefficients in the field of meromorphic functions of x, it therefore follows that $\tilde{B}_{k+1}$ is also. Since from Step k, $\{dx, \{d\tilde{y}_i^{(j)} | 1 \leqq i \leqq k, i \leqq j \leqq k\}\}$ is a linearly independent set over $\mathscr{K}$, it follows easily that $\tilde{B}_{k+1}$, when viewed as a rational function of $(u, \cdots, u^{(n-1)})$, has rank s_{k+1} over the field $\mathscr{K}$. Using (2.15), we show readily that

$$\begin{aligned} &\dim \operatorname{span} \{dx, \{d\tilde{y}_i^{(j)} | 1 \leqq i \leqq k, i \leqq j \leqq k\}\} \\ &\quad = \dim \operatorname{span} \{dx, \{\tilde{B}_\ell \, du^{(k+1-\ell)} | 1 \leqq \ell \leqq k+1\}\}. \end{aligned} \tag{2.16}$$

The dimension on the right-hand side of (2.16) is easily seen to be $n + s_1 + \cdots + s_{k+1}$, showing that, indeed, the vectors on the left-hand side of (2.16) are linearly independent since there are precisely $n + s_1 + \cdots + s_{k+1}$ elements; it only remains to show that they span $\mathscr{E}_{k+1}$. By its definition,

$$\mathscr{E}_{k+1} = \mathscr{E}_k + \operatorname{span} \{dy^{(k+1)}\}. \tag{2.17}$$

Using (2.15) once again, we can write this as follows:

$$\mathscr{E}_{k+1} = \mathscr{E}_k + \operatorname{span} \{d\tilde{y}_i^{(k+1)} | 1 \leqq i \leqq k+1\}, \tag{2.18}$$

which, coupled with the induction hypothesis, completes the proof. □

COROLLARY 2.4. *System* (1.1) *is left-invertible in the sense of Singh if and only if* $\dim \mathscr{E}_n - \dim \mathscr{E}_{n-1} = m$, *the number of scalar input components.*

2.3. The dynamic extension algorithm. The relationship between dynamic input-output decoupling and right invertibility has been clarified recently by Fliess [5], whose work has inspired several authors to develop concrete algorithms for the explicit construction of a dynamic compensator yielding a noninteractive system [3], [4]. Here we will give a simplified version of these algorithms, separating clearly their basic operations of differentiating the outputs, performing static-state feedback, and adding integrators on selected components of the input. This simplified version will still yield a decoupling compensator whenever one exists, but does introduce more integrators than the algorithms previously cited. We will show that it explicitly produces a basis for the chain of subspaces $\mathscr{E}_0 \subset \cdots \subset \mathscr{E}_n$. This shows yet another way in which the chain $\mathscr{E}_0 \subset \cdots \subset \mathscr{E}_n$ incorporates important structural information on a high level—in this case, that information pertraining to dynamic decoupling.

The dynamic extension algorithm is now presented.

Step 1. Let Σ_0 denote the system (1.1). Calculate

$$\dot{y} = \frac{\partial h(x)}{\partial x} [f(x) + g(x)u] \tag{2.19}$$

and write it as

$$\dot{y} = a_1(x) + b_1(x)u. \tag{2.20}$$

Define

$$\sigma_1 = \operatorname{rank} b_1(x), \tag{2.21}$$

where the rank is taken over the field of meromorphic functions of x. Permute, if necessary, the components of the output so that the first σ_1 rows of $b_1(x)$ are linearly independent. Decompose $\dot{y}$ as

$$\dot{y} = \begin{pmatrix} \dot{\bar{y}}_1 \\ \dot{\hat{y}}_1 \end{pmatrix} = \begin{pmatrix} \bar{a}_1(x) + \bar{b}_1(x)u \\ \hat{a}_1(x) + \hat{b}_1(x)u \end{pmatrix}. \tag{2.22}$$

Let

$$u = \alpha_1(x) + \beta_1(x)v_1 \tag{2.23}$$

be any static state feedback such that

(i) $\beta_1(x)$ is invertible over the field of meromorphic functions in x;

(ii) $\dot{\bar{y}}_1 = \bar{v}_1$.

Such a feedback always exists.

For the resulting closed-loop system, we can write

$$\dot{\bar{y}}_1 = \dot{\bar{y}}_1(x, \bar{v}_1), \tag{2.24}$$

since otherwise the rank of $\partial \dot{y}/\partial v_1$ would exceed σ_1. Moreover, $\dot{\bar{y}}_1$ is affine in $\bar{v}_1$. Now introduce a dynamic extension by

$$\dot{\bar{v}}_1 = \bar{u}_1 \tag{2.25a}$$

and rename the remaining components of v_1:

$$\hat{v}_1 = \hat{u}_1. \tag{2.25b}$$

Finally, let Σ_1 denote the system consisting of Σ_0, the static-state feedback (2.23), and the dynamic extension (2.25). Its state is given by $x_1 = \begin{pmatrix} x \\ \bar{v}_1 \end{pmatrix}$, its input is $u_1 = \begin{pmatrix} \bar{u}_1 \\ \hat{u}_1 \end{pmatrix}$, and the output remains $y = h(x)$. We will denote it as

$$\Sigma_1: \begin{aligned} \dot{x}_1 &= f_1(x_1) + g_1(x_1)u_1, \\ y &= h(x). \end{aligned} \tag{2.26}$$

Σ_1 is a dynamic extension of Σ_0 and has the property that $\dot{y} = \dot{y}(x_1)$ is affine in $\bar{v}_1$.

Step $k+1$. Suppose that in Step k the system Σ_k has been constructed such that

$$\Sigma_k: \begin{aligned} \dot{x}_k &= f_k(x_k) + g_k(x_k)u_k, \\ y &= h(x) \end{aligned} \tag{2.27}$$

and $y_k^{(k)} = y_k^{(k)}(x_k)$ is a rational function of $\bar{v}_1, \cdots, \bar{v}_k$ with coefficients in the field of meromorphic functions of x. Then, calculate

$$y^{(k+1)} = \frac{\partial y^{(k)}}{\partial x_k}[f_k(x_k) + g_k(x_k)u_k] \tag{2.28}$$

and rewrite (2.28) as

$$y^{(k+1)} = a_{k+1}(x_k) + b_{k+1}(x_k)u_k. \tag{2.29}$$

Define

$$\sigma_{k+1} = \operatorname{rank} b_{k+1}(x_k), \tag{2.30}$$

where the rank is over the field of rational functions of $\bar{v}_1, \cdots, \bar{v}_k$ with coefficients in the field of meromorphic functions of x. Permute, if necessary, the last $p-\sigma_k$ components of the output so that the first σ_{k+1} rows of $b_{k+1}(x_k)$ are linearly independent. Decompose $y^{(k+1)}$ as

$$y^{(k+1)} = \begin{pmatrix} \bar{y}_{k+1}^{(k+1)} \\ \hat{y}_{k+1}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \bar{a}_{k+1}(x_k) + \bar{b}_{k+1}(x_k)u_k \\ \hat{a}_{k+1}(x_k) + \hat{b}_{k+1}(x_k)u_k \end{pmatrix}. \tag{2.31}$$

Let

$$u_k = \alpha_{k+1}(x_k) + \beta_{k+1}(x_k)v_{k+1} \tag{2.32}$$

be any static-state feedback such that

(i) $\beta_{k+1}(x_k)$ is invertible over the field of rational functions of $\bar{v}_1, \cdots, \bar{v}_k$ with coefficients in the field of meromorphic functions of x.

(ii) $\bar{y}_{k+1}^{(k+1)} = \bar{v}_{k+1}$.

Such a feedback always exists.

For the resulting closed-loop system, we can always write

$$\hat{y}_{k+1}^{(k+1)} = \hat{y}_{k+1}^{(k+1)}(x_k, \bar{v}_{k+1}), \tag{2.33}$$

since otherwise the rank of $\partial y^{(k+1)}/\partial v_{k+1}$ would exceed σ_{k+1}. Moreover, $y^{(k+1)}$ is a rational function of $\bar{v}_1, \cdots, \bar{v}_{k+1}$. Now introduce a dynamic extension by

$$\dot{\bar{v}}_{k+1} = \bar{u}_{k+1} \tag{2.34a}$$

and rename the remaining components of v_{k+1}:

$$\hat{v}_{k+1} = \hat{u}_{k+1}. \tag{2.34b}$$

Finally, let Σ_{k+1} denote the system consisting of Σ_k, the static state feedback (2.32) and the dynamic extension (2.34). Its state is given by $x_{k+1} = \begin{pmatrix} x_k \\ \bar{v}_{k+1} \end{pmatrix}$, its input is $u_{k+1} = \begin{pmatrix} \bar{u}_{k+1} \\ \hat{u}_{k+1} \end{pmatrix}$, and the output remains $y = h(x)$. Σ_{k+1} is then a dynamic extension of Σ_k, has the property that $y^{(k+1)} = y^{(k+1)}(x_{k+1})$, and is a rational function of $\bar{v}_1, \cdots, \bar{v}_{k+1}$.

End of Step $k+1$.

We now have the following theorem.

THEOREM 2.5. *For each* $1 \leqq k \leqq n$:

(a) $\{dx, d\dot{\bar{y}}_1, \cdots, d\bar{y}_k^{(k)}\}$ *is a basis for* $\mathscr{E}_k$.

(b) $\dim \mathscr{E}_k = n + \sigma_1 + \cdots + \sigma_k$.

Proof. Part (b) is an immediate consequence of (a). From (2.33), one has that

$$\mathscr{E}_k = \operatorname{span}\{dx, d\dot{\hat{y}}_1, \cdots, d\bar{y}_k^{(k)}\}, \tag{2.35}$$

which by definition of $\bar{v}_j$ gives

$$\mathscr{E}_k = \operatorname{span}\{dx, d\bar{v}_1, d\bar{v}_2, \cdots, d\bar{v}_k\}. \tag{2.36}$$

Hence, it suffices to show that $\{dx, d\bar{v}_1, \cdots, d\bar{v}_k\}$ is a linearly independent set for each $1 \leqq k \leqq n$. This follows immediately from Lemma 2.6.

LEMMA 2.6. *For each* $1 \leqq k \leqq n-1$,

$$\begin{aligned} &\operatorname{span}\{dx, du, \cdots, du^{(n-1)}\} \\ &\quad = \operatorname{span}\{dx, d\bar{v}_1, \cdots, d\bar{v}_k, du_k, \cdots, du_k^{(n-1-k)}, d\hat{v}_k^{(n-k)}, \cdots, d\hat{v}_1^{(n-1)}\}, \end{aligned} \tag{2.37}$$

where all spans are with respect to the field $\mathscr{K}$.

Proof. Equations (2.23) and (2.25) yield

$$\begin{aligned}&\operatorname{span}\{dx, du, \cdots, du^{(n-1)}\}\\&\quad=\operatorname{span}\{dx, d\bar{v}_1, \{d\hat{v}_1, d\dot{\hat{v}}_1\}, \cdots, \{d\hat{v}_1^{(n-2)}, d\bar{v}_1^{(n-1)}\}, d\hat{v}_1^{(n-1)}\}\\&\quad=\operatorname{span}\{dx, d\bar{v}_1, du_1, \cdots, du_1^{(n-2)}, d\hat{v}_1^{(n-1)}\},\end{aligned}$$

which establishes (2.37) for $k=1$.

Now suppose that (2.37) holds at k. Then, in particular, $\{dx, d\bar{v}_1, \cdots, d\bar{v}_k\}$ is a linearly independent set. After we note that each component of $x_k=(x^T, \bar{v}_1^T, \cdots, \bar{v}_k^T)^T$ is a rational function of $(u, \cdots, u^{(n-1)})$, this gives us that b_{k+1} has full rank over $\mathscr{K}$. Hence, using (2.32) and (2.34), we have

$$\begin{aligned}&\operatorname{span}\{d\bar{x}_k, du_k, \cdots, du_k^{(n-1-k)}\}\\&\quad=\operatorname{span}\{d\bar{x}_k, d\bar{v}_{k+1}, \{d\hat{v}_{k+1}, d\dot{\hat{v}}_{k+1}\}, \cdots, \{d\hat{v}_{k+1}^{(n-k-2)}, d\bar{v}_{k+1}^{(n-1-k)}\}, d\hat{v}_{k+1}^{(n-1-k)}\}\\&\quad=\operatorname{span}\{d\bar{x}_k, d\bar{v}_{k+1}, du_{k+1}, \cdots, du_{k+1}^{(n-k-2)}, d\hat{v}_{k+1}^{(n-1-k)}\},\end{aligned}$$

which establishes (2.37) at $k+1$. □

2.4. Convergence of the chain $\mathscr{E}_0\subset\cdots\subset\mathscr{E}_n$. In a completely analogous fashion to the way the chain $\mathscr{E}_0\subset\cdots\subset\mathscr{E}_n$ is defined in (1.5), we could define an extended chain $\mathscr{E}_0\subset\cdots\subset\mathscr{E}_n\subset\cdots\subset\mathscr{E}_{n+k}$, for any finite integer $k\geqq 1$. The underlying field $\mathscr{K}_{n+k}$ would then consist of the set of rational functions of $(u, \cdots, u^{(n+k-1)})$ with coefficients that are meromorphic in x. We will show, however, that no new structural information is obtained by extending the original chain; hence, we are justified in terminating the chain at n, the dimension of the state space of (1.1).

THEOREM 2.7. *For all finite integers $k\geqq 1$:*

(a) $\{dx, d\dot{y}_1, \cdots, d\bar{y}_n^{(n)}, d\bar{y}_n^{(n+1)}, \cdots, d\bar{y}_n^{(n+\ell)}\}$ *is a basis for* $\mathscr{E}_{n+\ell}$, $1\leqq\ell\leqq k$.

(b) $\dim\mathscr{E}_{n+\ell}=\dim\mathscr{E}_n+\ell\cdot\rho^*$, $1\leqq\ell\leqq k$, *where* $\rho^*=\dim\mathscr{E}_n-\dim\mathscr{E}_{n-1}$.

The proof will be broken down into two lemmas. The functions $(\bar{y}_i^T, \hat{y}_i^T)=y^T$ continue to denote the blocks of the output constructed in the ith step of the dynamic extension algorithm; $\bar{y}_i^{(j)}$, $\hat{y}_i^{(j)}$ denote their jth-order derivatives along the dynamics of the system.

LEMMA 2.8. *There exists an integer $1\leqq N\leqq n$ such that*

$$d\hat{y}_n^{(N)}\in\operatorname{span}_{\mathscr{K}}\{d\hat{y}_n, \cdots, d\hat{y}_n^{(N-1)}, d\dot{\hat{y}}_1, \cdots, d\bar{y}_n^{(N)}\}. \tag{2.38}$$

Proof. From the dynamic extension algorithm, we have that $\hat{y}_k^{(k)}=\hat{y}_k^{(k)}(x, \dot{\hat{y}}_1, \cdots, \bar{y}_k^{(k)})$. Hence, (2.38) holds if

$$\frac{\partial\hat{y}_n^{(N)}}{\partial x}dx\in\operatorname{span}_{\mathscr{K}}\left\{\frac{\partial\hat{y}_n}{\partial x}dx, \cdots, \frac{\partial\hat{y}_n^{(N-1)}}{\partial x}dx\right\}. \tag{2.39}$$

Since the right-hand side of (2.39) can be at most n-dimensional, there must exist $1\leqq N\leqq n$ such that (2.39) holds. Hence, the result is proved. □

For each integer j, $N\leqq j\leqq n+k$, define the field $\mathscr{K}_j$ to be the set of rational functions of $(u, \cdots, u^{(j-1)})$ with coefficients that are meromorphic in x.

LEMMA 2.9. *With N as in Lemma 2.8 and for all integers j, $N\leqq j\leqq n+k$,*

$$d\hat{y}_n^{(j)}\in\operatorname{span}_{\mathscr{K}_j}\{d\hat{y}_n, \cdots, d\hat{y}_n^{(N-1)}, d\dot{\hat{y}}_1, \cdots, d\bar{y}_j^{(j)}\}, \tag{2.40}$$

where, for $j\geqq n$, $\bar{y}_j=\bar{y}_n$.

Proof. Since all the functions appearing in (2.38) are rational functions of $(u, \cdots, u^{(N-1)})$, (2.38) also holds with $\mathscr{K}$ replaced by $\mathscr{K}_N$. Hence there exist coefficients $\alpha_i, \beta_i \in \mathscr{K}_N$ such that

$$d\hat{y}_n^{(N)} = \sum_{i=0}^{N-1} \alpha_i \, d\hat{y}_n^{(i)} + \sum_{i=1}^{N} \beta_i \, d\bar{y}_i^{(i)}. \tag{2.41}$$

Hence,

$$d\hat{y}_n^{(N+1)} = \sum_{i=0}^{N-1} (\dot{\alpha}_i \, d\hat{y}_n^{(i)} + \alpha_i \, d\hat{y}_n^{(i+1)}) + \sum_{i=1}^{N} (\dot{\beta}_i \, d\bar{y}_i^{(i)} + \beta_i \, d\bar{y}_i^{(i+1)}). \tag{2.42}$$

Combining (2.41) with the fact that $\dot{\alpha}_i$, $\dot{\beta}_i$ are in $\mathscr{K}_{N+1}$, we easily see that (2.42) establishes (2.40) for $j = N+1$. In a similar manner we complete the proof of Lemma 2.9. □

To complete the proof of Theorem 2.7, note that Lemma 2.9 establishes that $\mathscr{E}_{n+\ell}$, $1 \leqq \ell \leqq k$ is indeed spanned by the vectors given in (a). We prove that they are independent using the same reasoning employed in the proof of Theorem 2.3. Part (b) follows from (a) by counting the number of basis elements.

Remark 2.10. If we introduce the finite chain of subspaces $\mathscr{F}_0 \subset \cdots \subset \mathscr{F}_n$ of $\mathscr{E}$ by $\mathscr{F}_k = \operatorname{span}\{dy, d\dot{y}, \cdots, dy^{(k)}\}$, then we can show that $\rho^* = \dim \mathscr{F}_n - \dim \mathscr{F}_{n-1}$. However, the inversion algorithm and the dynamic extension algorithm do not calculate bases for this chain.

2.5. Differential output rank. In 1985, Fliess introduced a new approach, centered around differential algebra, to the analysis of nonlinear systems, [5], [16], the proper formalism being perhaps field theoretic [17]. He was the first to define clearly and precisely the fundamental notion of the *rank* of a nonlinear system; he accomplished this by considering the output components to be dependent if they satisfied a nontrivial (nonlinear) differential equation. Fliess' notion of rank generalized to nonlinear systems the usual notion of the rank of a transfer matrix of a linear system, and played the same important role in leading to basic definitions of right invertibility and left invertibility, and important new results on dynamic feedback [5], [16]. All previous attempts in this regard lacked the power and elegance of the differential algebraic approach, being mainly based on algorithmic considerations.

Additional insights on the role of differential algebra in system theory have been contributed by Pommaret [18].

In this section, we will show that the integer $\rho^* = \dim \mathscr{E}_n - \dim \mathscr{E}_{n-1}$ of Theorem 2.7 is actually the differential output rank for those systems that meet the requirements both of § 1 and of the differential algebraic approach. More precisely, we suppose that the system (1.1) satisfies the following assumptions:

(A1) f, g and h are meromorphic;

(A2) f, g and h are differentially algebraic (i.e., elementary transcendental) functions of their arguments [5], [19], [20];

(A3) The set $\mathbb{R}\langle y \rangle$ of all rational functions of $y_i^{(\ell)}$, $1 \leqq i \leqq p$, $\ell \geqq 0$, with coefficients in the field $\mathbb{R}$ is a differential field.

Here, the $y_i^{(\ell)}$ are defined as in (1.4). (It is apparently unknown whether (A1) implies (A3).)

To aid the reader, a few notions from differential algebra are briefly recalled. A finite set of elements $\zeta_1, \cdots, \zeta_k$ of $\mathbb{R}\langle y \rangle$ *is differentially algebraically dependent* if there exists a nonzero differential polynomial $\mathbf{P}$, with coefficients in $\mathbb{R}$, such that

$\mathbf{P}(\zeta_1, \cdots, \zeta_k) = 0$; that is, a nonzero polynomial in $\zeta_1, \cdots, \zeta_k$ and a finite number of their time derivatives. The *differential output rank*, denoted $d^0(\Sigma)$, is the maximum number of differentially algebraically independent elements of $\mathbb{R}\langle y\rangle$. This number is well defined [21].

Our main result relating the $d^0(\Sigma)$ to the chain $\mathscr{E}_0 \subset \cdots \subset \mathscr{E}_n$ will be an easy consequence of the following properties of the differential rank.

LEMMA 2.11 [5]-[7], [16]. (a) $d^0(\Sigma) \leqq \min\{m, p\}$.

(b) *If Σ' is obtained from Σ by applying an invertible static state feedback, whose components are differentially algebraic functions, then $d^0(\Sigma') = d^0(\Sigma)$.*

(c) *If Σ_e is obtained by adding a finite number k_i of integrators on each input channel i of Σ, then $d^0(\Sigma_e) = d^0(\Sigma)$.*

(d) *If for some $1 \leqq i \leqq m$, $\partial y^{(k)}/\partial u_i^{(\ell)} = 0$ for all $k \geqq 0$, $\ell \geqq 0$, then $d^0(\Sigma) \leqq \min\{m-1, p\}$.*

(e) *If for some finite integer k, $y^{(k)} = y^{(k)}(x, u)$, and rank $\partial y^{(k)}/\partial u = r$, then $d^0(\Sigma) \geqq r$.*

Proof. All of these points are essentially contained in [5]-[7], [16]. Part (a) follows from the definition of the differential output rank. Part (b) is true because $d^0(\Sigma') \leqq d^0(\Sigma)$ for any static-state feedback. Thus invertibility gives equality, since the inverse, being a well-defined (differentially algebraic) static-state feedback, yields $d^0(\Sigma) \leqq d^0(\Sigma')$. To prove (c), first note that

$$m = \text{diff. tr. } d^0 \frac{\mathbb{R}\langle u, y\rangle}{\mathbb{R}} = \text{diff. tr. } d^0 \frac{\mathbb{R}\langle u, y\rangle}{\mathbb{R}\langle y\rangle} + \text{diff. tr. } d^0 \frac{\mathbb{R}\langle y\rangle}{\mathbb{R}}, \tag{2.43}$$

which yields

$$d^0(\Sigma) = m - \text{diff. tr. } d^0 \frac{\mathbb{R}\langle u, y\rangle}{\mathbb{R}\langle y\rangle}. \tag{2.44}$$

Similarly,

$$d^0(\Sigma_e) = m - \text{diff. tr. } d^0 \frac{\mathbb{R}\langle v, y\rangle}{\mathbb{R}\langle y\rangle}. \tag{2.45}$$

This establishes the result because the right-hand sides of (2.44) and (2.45) are equal, since v constitutes a differential transcendence basis for $\mathbb{R}\langle u\rangle/\mathbb{R}$. Statement (d) follows from (a) because y is differentially algebraic over $\mathbb{R}\langle u_1, \cdots, u_{i-1}, u_{i+1}, \cdots, u_m\rangle$. Finally, (e) implies that

$$\dim_{\mathscr{K}} \{d\dot{y}, \cdots, dy^{(k)}\} \geqq r, \tag{2.46}$$

which yields

$$\dim_{\mathbb{L}} \{d\dot{y}, \cdots, dy^{(k)}\} \geqq r, \tag{2.47}$$

where $\mathbb{L}$ is the field of rational functions in the indeterminates $\{y, \dot{y}, \cdots, y^{(k)}\}$ with coefficients in $\mathbb{R}$; note that $\mathbb{L} \subset \mathscr{K}$. Hence, by the results in [22], $d^0(\Sigma) \geq r$. □

THEOREM 2.12. *Suppose that system* (1.1) *satisfies Assumptions* (A1)-(A3). Then

$$d^0(\Sigma) = \dim \mathscr{E}_n - \dim \mathscr{E}_{n-1}. \tag{2.48}$$

Proof. Consider the extended system Σ_n constructed at the nth step of the dynamic extension algorithm. From (2.33), Σ_n has the property that rank $\partial y^{(n)}/\partial v_n = \sigma_n$. Hence, by (e) of Lemma 2.11, $d^0(\Sigma_n) \geqq \sigma_n$. By Theorem 2.7, $\sigma_{n+\ell} = \sigma_n$ for all $\ell \geqq 0$; therefore, y satisfies $\partial y^{(k)}/\partial \hat{v}_n^{(\ell)} = 0$ for all $k \geqq 0$, $\ell \geqq 0$. Thus, by (d) of Lemma 2.11, $d^0(\Sigma_n) \leqq \sigma_n$.

Hence, $d^0(\Sigma_n) = \sigma_n$. However, (b) and (c) of Lemma 2.11 then yield that $d^0(\Sigma) = d^0(\Sigma_n) = \sigma_n$, because Σ_n is constructed from Σ by successively applying invertible static feedbacks and finite strings of integrators. Finally, Theorem 2.5 yields $\sigma_n = \dim \mathscr{E}_n - \dim \mathscr{E}_{n-1}$. □

3. Structure at infinity and block decoupling. In this section, the framework of §§ 1 and 2 will be used to give an intrinsic "algebraic" definition of some important integer invariants associated with a nonlinear system, namely the so-called structure at infinity. An alternate approach to defining a structure at infinity has already been carried out in [8], using differential geometric techniques. When specialized to the class of linear systems, the geometric definition and the algebraic approach below both agree with the usual linear notion of the structure at infinity [23]. For general nonlinear systems, both are invariant under regular *static* state feedbacks. However, in contrast to the geometric approach, the algebraically based definition enjoys some additional properties that make it seem closer to the linear situation.

DEFINITION 3.1 [9]. The *number σ_k of zeros at infinity of order less than or equal to k, $k \geqq 1$,* is $\sigma_k = \dim \mathscr{E}_k - \dim \mathscr{E}_{k-1}$. The *structure at infinity* is given by the list $\{\sigma_1, \cdots, \sigma_n\}$.

Note that the total number of zeros at infinity σ_n corresponds precisely to the rank ρ^* of the system (1.1).

The notion of a *regular dynamic feedback* is introduced next. Note that when $q = 0$ in the following definition, we recover the usual definition of a regular static feedback.

DEFINITION 3.2 [24]. The compensator

$$\dot{z} = F(x, z) + G(x, z)v, \qquad u = \alpha(x, z) + \beta(x, z)v, \tag{3.1}$$

where F, G, α, and β are meromorphic, $v \in \mathbb{R}^m$, and $z \in \mathbb{R}^q$ for a given integer q, is said to be *regular* if the composite system (1.1a) and (3.1), with v as the input and u as the output, has rank equal to m.

The structure at infinity of Definition 3.1 enjoys the following properties.

LEMMA 3.3. (a) *The rank ρ^* of the system* (1.1) *is equal to σ_n, the total number of zeros at infinity.*

(b) $\sigma_n \leqq \min\{m, p\}$, *the number of input and output components, respectively.*

(c) *The total number of zeros at infinity is invariant under regular dynamic feedback.*

Proof. Properties (a) and (b) are immediate from the results of § 2. To prove (c), first introduce the field $\mathscr{K}_e$ consisting of the set of rational functions of $(v, \cdots, v^{(n+q-1)})$ with coefficients which are meromorphic in x and z. Assume that for a given l, $0 \leqq l \leqq n+q-1$, there exists i, $1 \leqq i \leqq m$, such that

$$du_i^{(l)} \in \operatorname{span}_{\mathscr{K}_e}\{dx, du, \cdots, du^{(l-1)}, du_j^{(l)}, j \neq i\}.$$

Then, following the calculations involved in the proof of Lemma 2.9, we see that

$$du_i^{(k)} \in \operatorname{span}_{\mathscr{K}_e}\{dx, du, \cdots, du^{(k-1)}, du_j^{(k)}, j \neq i\} \quad \forall k \geqq l,$$

which contradicts the regularity assumption of the dynamic compensator. Hence, for every l, $\{dx, du, \cdots, du^{(l)}\}$ is a linearly independent set. Let $y_e = h(x)$ denote the output of the composite system Σ_e consisting of (1.1) and (3.1), and define

$$\mathscr{G}_k = \operatorname{span}_{\mathscr{K}_e}\{dx, dy_e, \cdots, dy_e^{(k)}\}.$$

By the chain rule,

$$\frac{\partial(x, y_e, \cdots, y_e^{(k)})}{\partial(x, z, v, \cdots, v^{(k-1)})} = \frac{\partial(x, y, \cdots, y^{(k)})}{\partial(x, u, \cdots, u^{(k-1)})} \frac{\partial(x, u, \cdots, u^{(k-1)})}{\partial(x, z, v, \cdots, v^{(k-1)})}. \tag{3.2}$$

Since $\{dx, du, \cdots, du^{(k-1)}\}$ are independent, (3.2) yields that

$$\dim \mathcal{G}_k = \operatorname{rank} \frac{\partial(x, y, \cdots, y^{(k)})}{\partial(x, u, \cdots, u^{(k-1)})},$$

so that

$$\dim \mathcal{G}_k = \dim \mathcal{E}_k. \tag{3.3}$$

Finally, let ρ_e^* denote the rank Σ_e. Then, following the reasoning employed in Lemma 2.8, we show that $\rho_e^* = \dim \mathcal{G}_{n+q} - \dim \mathcal{G}_{n+q-1}$, which yields the result in view of (3.3). □

Now consider a system whose outputs have been grouped into blocks:

$$\dot{x} = f(x) + g(x)u, \tag{1.1a}$$

$$y_i = h_i(x), \qquad 1 \leqq i \leqq \mu, \tag{3.4}$$

where $y_i \in \mathbb{R}^{p_i}$ and h_i is a meromorphic function of x. The system is said to be decoupled with respect to a given partition $u = (u_1^T, \cdots, u_\mu^T)^T$ of the input if u_i affects only y_i, $1 \leqq i \leqq \mu$; that is,

$$dy_i^{(k)} \in \operatorname{span}\{dx, du_i, \cdots, du_i^{(n-1)}\} \tag{3.5}$$

for $0 \leqq k \leqq n$. The decoupling problem is to find, if possible, a regular dynamic compensator and a partition of the new reference input such that the resulting closed-loop system is decoupled. If the compensator has dimension zero, the solution is said to be static; otherwise it is dynamic.

Using (3.5), we immediately obtain the following necessary condition for regular static block decoupling.

PROPOSITION 3.4 (see also [8]). *If the system* (1.1a), (3.4) *with block-partitioned outputs can be decoupled with a regular static state variable feedback, then*

$$\sigma_k = \sum_{i=1}^{\mu} \sigma_k^i, \tag{3.6}$$

where $\{\sigma_1^i, \cdots, \sigma_n^i\}$ *is the structure at infinity of the ith subsystem consisting of the dynamics* (1.1a) *and the output* y_i.

When the outputs y_i are scalar-valued, it is known that (3.6) is also sufficient (see, for example, [9], [25]). This condition is also known to be sufficient for general vector-valued outputs if we restrict our attention to the class of linear systems [26]. Indeed, (3.6) implies that we can perform the inversion algorithm for the overall system (1.1a), (3.4) by applying it to each of the individual subsystems y_i, for $1 \leqq i \leqq \mu$. Then, specializing equation (2.10) to the nth step of the algorithm and invoking linearity, we see that the $\tilde{b}_i$'s are constants, and the $\tilde{a}_i$'s are linear functions of x and various derivatives of y_i. Decoupling is accomplished by cancelling the dependence on x, and diagonalizing the matrix multiplying the inputs, via a static feedback.

The fact that condition (3.6) is not sufficient for general nonlinear systems is illustrated by the following example.

Example 3.5. Consider the system

$$\dot{x} = (u_1, x_4 + x_5 u_1, u_2, x_3 u_1, u_3)^T,$$

$$y_1 = (x_1, x_2)^T, \qquad y_2 = x_3.$$

We easily calculate that $\{\sigma_1, \sigma_2\} = \{2, 3\}$, and for the two subsystems $\{\sigma_1^1, \sigma_2^1\} = \{1, 2\}$ and $\{\sigma_1^2\} = \{1\}$. Thus, (3.6) is fulfilled. Nevertheless, a straightforward application of the results of Nijmeijer and Schumacher [8] shows that the system cannot be decoupled by any regular static-state feedback.

This example underlines the importance of the differential geometric approach in general, and the "geometric" structure at infinity in particular, for the study of static state feedback control problems, since the equivalent of (3.6) for the geometric structure at infinity constitutes a local necessary and sufficient condition (at regular points) [8]. On the other hand, algebraic methods seem to be better when we are studying dynamic feedback problems [3], [4], [27]. In this spirit, we have the following result that does not hold with the geometric version of the structure at infinity because it fails [10] the properties of Lemma 3.3.

THEOREM 3.6. *The system* (1.1a), (3.4) *can be decoupled with regular dynamic state feedback if and only if*

$$\rho = \sum_{i=1}^{\mu} \rho_i, \tag{3.7}$$

where ρ_i *denotes the rank of the subsystem* (1.1a) *with output* y_i.

Proof. Because the necessity is clear, only the sufficiency will be shown. For each block of outputs y_i, permute if necessary the components in such a way that $y_i = (\bar{y}_i^T, \hat{y}_i^T)^T$, $\dim \bar{y}_i = \rho_i$, and on defining

$$\mathscr{E}_k^i = \operatorname{span}\{dx, d\dot{y}_i, \cdots, dy_i^{(k)}\},$$

we have

$$\mathscr{E}_k^i = \operatorname{span}\{dx, d\dot{\bar{y}}_i, \cdots, d\bar{y}_i^{(k)}\}$$

for all k, $1 \leqq k \leqq n$. This can always be accomplished with the help of the inversion algorithm of § 2.2. Then,

$$d\hat{y}_i^{(k)} \in \operatorname{span}\{dx, d\dot{\bar{y}}_i, \cdots, d\bar{y}_i^{(k)}\}$$

and the rank of the subsystem (1.1a) with output $\bar{y}_i$ is equal to ρ_i. Let $\bar{\Sigma}$ denote the subsystem whose output is given by $(\bar{y}_1, \cdots, \bar{y}_\mu)$. We conclude that the rank of $\bar{\Sigma}$ is ρ and the original system (1.1a), (3.4) can be decoupled if and only if $\bar{\Sigma}$ can be. It follows from (3.7) that the rank of $\bar{\Sigma}$ equals the number of its scalar output components; thus, $\bar{\Sigma}$ can be row-wise decoupled [25].

4. Conclusions. In the recent literature, there have been many attempts to extend concepts and tools from the linear setting to the class of nonlinear systems. For example, Nijmeijer extends a definition of right invertibility based on the consideration of the sequence of Toeplitz matrices associated to a linear system. Singh extends the notion of left invertibility based on the input elimination idea of Silverman's algorithm. Descusse and Moog and Nijmeijer and Respondek use the idea of delaying the inputs via the addition of integrators, as does Wang [28], to achieve dynamic decoupling of nonlinear systems. Fliess uses differential algebra to extend the notion of the rank of a transfer matrix, so as to synthesize left invertibility, right invertibility, and dynamic decoupling.

In this paper, we have shown that all of the above extensions can be unified by the study of a particular chain of subspaces naturally associated to the output of a system. As simple corollaries of our analysis, we obtain that right invertibility in the sense of Nijmeijer is the same as in that of Fliess, which in turn is the same as $\rho^* = p$, the number of scalar output components. Left invertibility in the sense of Singh is the

same as in that of Fliess, which is the same as $\rho^* = m$, the number of scalar input components. Moreover, the inversion algorithm has a natural interpretation as a procedure for constructing a basis adapted to the chain of subspaces $\mathscr{E}_0 \subset \cdots \subset \mathscr{E}_n$. In a similar vein, the dynamic extension algorithm, which lies at the heart of dynamic decoupling, also constructs a basis adapted to the chain $\mathscr{E}_0 \subset \cdots \subset \mathscr{E}_n$. The main difficulty in comparing the various algorithms was that each was working over its own unique field. This was overcome by relating each of the fields to a common larger field. In this way, the equivalence of four topics, that previously had only been studied separately, was established.

The search for a nonlinear version of the structure at infinity of a linear system has incited much effort on the part of many researchers [2], [7]-[9], [29], with the goal of finding an appropriate tool for solving such classical synthesis problems as noninteracting control and model matching. One of the first efforts in this regard was perceived to possess certain deficiencies [10], because a system could have more zeros at infinity than input or output components, and also because the number of zeros at infinity could be altered by the addition of integrators on the input channels.

Section 3 takes an algebraic approach to defining a nonlinear structure at infinity [9]. Its number of zeros at infinity is always less than or equal to the number of inputs or outputs, and is invariant under the action of regular dynamic feedback. However, the deficiency of this generalization of the structure at infinity is that it cannot properly address the static block noninteracting control problem. Hence the "right" approach, if it exists [30], is yet to be discovered.

REFERENCES

[1] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, AC-26, (1981), pp. 595-598.

[2] H. NIJMEIJER, *Right-invertibility for a class of nonlinear control systems: a geometric approach*, Systems Control Lett., 7 (1986), pp. 125-132.

[3] J. DESCUSSE AND C. H. MOOG, *Dynamic decoupling for right-invertible nonlinear systems*, System Control Lett., 8 (1987), pp. 345-349.

[4] H. NIJMEIJER AND W. RESPONDEK, *Decoupling via dynamic compensation for nonlinear control systems*, in Proc. 25th Conference on Decision and Control, Athens, 1986, pp. 192-197.

[5] M. FLIESS, *A new approach to the noninteracting control problem in nonlinear systems theory*, in Proc. 23rd Allerton Conference, University of Illinois, Monticello, IL, 1985, pp. 123-129.

[6] ———, *A note on the invertibility of nonlinear input-output differential systems*, Systems Control Lett., 8 (1986), pp. 147-151.

[7] ———, *A new approach to the structure at infinity of nonlinear systems*, Systems Control Lett., 7 (1986), pp. 419-421.

[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] C. H. MOOG, *Nonlinear decoupling and structure at infinity*, Math. Control Signals Systems, 1 (1988), pp. 257-268.

[10] A. ISIDORI, *Control of nonlinear systems via dynamic state-feedback*, in Algebraic and Geometric Methods in Nonlinear Control Theory, Proc. Conference Paris, 1985, M. Fliess and M. Hazewinkel, eds., Reidel, Dordrecht, the Netherlands, 1986, pp. 121-145.

[11] J. W. GRIZZLE, M. D. DI BENEDETTO, AND C. H. MOOG, *Computing the differential output rank of a nonlinear system*, in Proc. 26th Conference on Decision and Control, Los Angeles, December 1987, pp. 142-145.

[12] R. ABRAHAM AND J. E. MARSDEN, *Foundations of Mechanics*, Second edtion, Benjamin-Cummings, Reading, May, 1980.

[13] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 270-276.

[14] R. M. HIRSCHORN, *Invertibility of multivariable nonlinear control systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 855-865.

[15] A. ISIDORI AND C. H. MOOG, *On the equivalent of the notion of transmission zeros*, in Modelling and Adaptive Control, Proc. Mathematical Proc. International Institute of Applied Systems Analysis Conference, Sopron, 1986, C. I. Byrnes and A. Kurszanski, eds., Lecture Notes in Control Information Science, 105, Springer-Verlag, Berlin, New York, 1988.

[16] M. FLIESS, *Some remarks on nonlinear invertibility and dynamic state-feedback*, in Theory and Applications of Nonlinear Control Systems, Mathematical Theory of Networks and Systems 85, Stockholm, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, New York, 1986, pp. 115-121.

[17] ———, *Nonlinear control theory and differential algebra*, in Modelling and Adaptive Control, Proc. International Institute of Applied Systems Analysis Conference, Sopron, 1986, C. I. Byrnes and A. Kurszanski, eds., Lecture Notes in Control and Information Science, 105, Springer-Verlag, Berlin, New York, 1988.

[18] J. F. POMMARET, *Géométrie différentielle algébrique et théorie du contrôle*, C.R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 547-550.

[19] L. A. RUBEL AND M. F. SINGER, *A differentially algebraic elimination theorem with application to analog computability in the calculus of variations*, Proc. Amer. Math. Soc., 94 (1985), pp. 653-658.

[20] J. F. POMMARET, *Differential Galois Theory*, Gordon and Breach, New York, 1983.

[21] E. R. KOLCHIN, *Differential Algebra and Algebraic Groups*, Academic Press, New York, 1973.

[22] J. JOHNSON, *Kähler differentials and differential algebra*, Ann. of Math., 89 (1969), pp. 92-98.

[23] L. M. SILVERMAN AND A. KITAPÇI, *System structure at infinity*, Systems Control Lett., 3 (1983), pp. 123-131.

[24] C. H. MOOG, *Note on the left-invertibility of nonlinear systems*, in Proc. Mathematical Theory of Networks and Systems 87, Phoenix, AZ, to appear.

[25] 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.

[26] J. DESCUSSE, J. F. LAFAY, AND M. MALABRE, *On the structure at infinity of linear block-decouplable systems: the general case*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 1115-1118.

[27] M. FLIESS, *Vers une nouvelle théorie du bouclage dynamique sur la sortie des systèmes non linéaires*, in Analysis and Optimization of Systems, Lecture Notes. Contr. Inform. Sci. 83, Springer-Verlag, Berlin, 1986, pp. 293-299.

[28] S. H. WANG, *Design of precompensator for decoupling problem*, Electronics Letters, 6 (1970), pp. 739-741.

[29] G. CONTE, M. D. DI BENEDETTO, A. ISIDORI, AND A. M. PERDON, *An input-output characterization of the structure at infinity for a nonlinear system*, in Theory and Applications of Nonlinear Control Systems, Mathematical Theory of Networks and Systems 85, Stockholm, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, 1986, pp. 301-312.

[30] M. FLIESS, Personal communication, 1986.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 673-696, May 1989

ANALYSE ASYMPTOTIQUE ET PROBLEME HOMOGENEISE EN CONTROLE OPTIMAL AVEC VIBRATIONS RAPIDES*

SHI-GE PENG†

Abstract. Le but de cet article est de discuter un problème de contrôle optimal avec vibrations rapides. Le modèle mathématique est le suivant: pour $\varepsilon > 0$, assez petit, on considère

$$(*)_1 \qquad \frac{dx^\varepsilon}{dt} = g\left(\frac{t}{\varepsilon}, x^\varepsilon(t), v(t)\right), \qquad x^\varepsilon(0) = x_0$$

où $v(\cdot) \in L^2(0, T; \mathbb{R}^k)$ est la fonction de contrôle. $x^\varepsilon(\cdot) \in H^1(0, T; \mathbb{R}^n)$ est la fonction d'état. $g(\tau, x, v)$ est 1-périodique par rapport à $\tau \in \mathbb{R}$. On cherche à minimiser la fonction coût

$$(*)_2 \qquad \int_0^T l\left(\frac{t}{\varepsilon}, x^\varepsilon(t), v(t)\right) dt$$

où $l(\tau, x, v)$ est aussi 1-périodique par rapport à τ.

On s'intéresse tout d'abord au comportement du système quand $\varepsilon \to 0$. On pourrait peut-être croire que ce problème va tendre vers le problème "moyen"

$$\frac{dx}{dt} = \bar{g}(x(t), v(t)), \qquad \min_{v(\cdot)} \int_0^T \bar{l}(x(t), v(t))\, dt$$

où

$$\bar{g}(x, v) = \int_0^1 g(\tau, x, v)\, d\tau, \qquad \bar{l}(x, v) = \int_0^1 l(\tau, x, v)\, d\tau.$$

Mais c'est faux en général. En réalité, le problème de limite de $(*)$ sera

$$(**) \qquad \frac{dx}{dt} = \int_0^1 g(\tau, x(t), v(t, \tau))\, d\tau, \qquad \min_{v(\cdot,\cdot)} \int_0^T \int_0^1 l(\tau, x(t), v(t, \tau))\, d\tau\, dt$$

où

$$v(\cdot, \cdot) \in L^2([0, T] \times [0, 1]; \mathbb{R}^k).$$

C'est donc une sorte de généralisation du problème de contrôle optimal usuel. On va l'appeler le problème homogénéisé du problème $(*)$. Mais on va voir que l'on peut caractériser le problème $(**)$ de façon parallèle au problème classique. On aura le principe de Pontryagin et la méthode de programmation dynamique. On va également établir le développement asymptotique et la convergence des problèmes $(*)$.

Key words. optimal control, homogenization, perturbations, periodic

AMS(MOS) subject classifications. 49, 93

1. Introduction.

Soit

$$(1) \qquad g(\tau, x, v) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}^n, \quad l(\tau, x, v) : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^k \to \mathbb{R}, \quad h(x) : \mathbb{R}^n \to \mathbb{R}$$

trois fonctions mesurables. On suppose que

$$(2) \qquad \begin{aligned} &g \text{ et } l \text{ sont 1-périodiques par rapport à } \tau, \\ &g(\tau + 1, x, v) = g(\tau, x, v), \\ &l(\tau + 1, x, v) = l(\tau, x, v) \quad \forall \tau, x, v. \end{aligned}$$

* Received by the editors December 29, 1986; accepted for publication (in revised form) September 14, 1988. This work was finished during the author's stay at the U.E.R. de Mathématiques, Université de Provence (Marseille, France) and was supported in part by the Institut National de Recherche en Informatique et en Automatique (INRIA), France.

† Mathematics Department, Shandong University, Jinan, People's Republic of China.

On suppose également que g est Lipschitzien par rapport à x. Pour $\varepsilon > 0$ assez petit, on va chercher le contrôle optimal du problème

$$\begin{aligned} &\frac{dx^\varepsilon}{dt} = g\left(\frac{t}{\varepsilon}, x^\varepsilon(t), v(t)\right), \qquad x^\varepsilon(0) = x_0, \\ &\min_{v(\cdot)} \left[\int_0^T l\left(\frac{t}{\varepsilon}, x^\varepsilon(t), v(t)\right) dt + h(x^\varepsilon(t))\right]. \end{aligned} \tag{3}$$

Formellement on aura les conditions nécessaires d'optimalité (principe de Pontryagin [9])

$$\begin{aligned} &\frac{dx^\varepsilon}{dt} = g\left(\frac{t}{\varepsilon}, x^\varepsilon(t), u^\varepsilon(t)\right), \\ &-\frac{dp^\varepsilon}{dt} = H_x\left(\frac{t}{\varepsilon}, x^\varepsilon(t), u^\varepsilon(t), p^\varepsilon(t)\right), \\ &H_v\left(\frac{t}{\varepsilon}, x^\varepsilon(t), u^\varepsilon(t), p^\varepsilon(t)\right) = 0 \end{aligned} \tag{4}$$

où $H(\tau, x, v, p) = p^* g(\tau, x, v) + l(\tau, x, v)$ est l'Hamiltonien.

On va étudier le comportement du problème de contrôle optimal quand $\varepsilon \to 0$. On va s'apercevoir que ce problème ne tend pas vers le "averaging problem"

$$\frac{dx}{dt} = \bar{g}(x(t), v(t)), \qquad \min_{v(\cdot)} \int_0^T \bar{l}(x(t), v(t))\, dt \tag{5}$$

où

$$\bar{g}(x, v) = \int_0^1 g(\tau, x, v)\, d\tau, \qquad \bar{l}(x, v) = \int_0^1 l(\tau, x, v)\, d\tau.$$

Le problème de limite est en fait une sorte de généralisation du problème de contrôle optimal où l'équation d'état s'écrit:

$$\frac{dx}{dt} = \int_0^1 g(\tau, x(t), v(t, \tau))\, d\tau, \qquad x(0) = x_0 \tag{6}$$

et le contrôle est une fonction de deux variables

$$v : [0, T] \times [0, 1] \to \mathbb{R}^k \quad \text{mesurable.}$$

Ici on souligne que la fonction d'état $x(t)$ est seulement une fonction de $t \in [0, T]$. De la même façon, la fonctionnelle à minimiser est

$$J^0(v(\cdot, \cdot)) = \int_0^1 \int_0^T l(\tau, x(t), v(t, \tau))\, d\tau\, dt + h(x(T)). \tag{7}$$

Ce phénomène peut se comprendre par le développement asymptotique des conditions de Pontryagin (4). On peut écrire formellement les développements de $x^\varepsilon(t)$, $u^\varepsilon(t)$, $p^\varepsilon(t)$ par la technique de changement de l'echelle du temps (voir [5], [6])

$$\begin{aligned} x^\varepsilon(t) &= x(t) + \varepsilon x_1\left(t, \frac{t}{\varepsilon}\right) + \cdots, \\ p^\varepsilon(t) &= p(t) + \varepsilon p_1\left(t, \frac{t}{\varepsilon}\right) + \cdots, \\ u^\varepsilon(t) &= u\left(t, \frac{t}{\varepsilon}\right) + \varepsilon u_1\left(t, \frac{t}{\varepsilon}\right) + \cdots \end{aligned}$$

où $u(t, \tau)$, $x_1(t, \tau)$, $p_1(t, \tau)$ sont 1-périodiques par rapport à τ. La raison pour laquelle le terme $u(t, (t/\varepsilon))$ contient déjà les vibrations rapides est la suivante: le contrôle réagit aux vibrations rapides du système beaucoup plus rapidement que la fonction d'état. En remplaçant x^ε, u^ε, p^ε par leur développement dans (4), pour l'ordre ε^0, on a $(\tau = t/\varepsilon)$

$$\begin{aligned} &\frac{\partial x_1}{\partial \tau} + \frac{dx}{dt} = g(\tau, x(t), u(t, \tau)), \\ &-\frac{\partial p_1}{\partial \tau} - \frac{dp}{dt} = H_x(\tau, x(t), u(t, \tau), p(t)), \\ &H_v(\tau, x(t), u(t, \tau), p(t)) = 0. \end{aligned} \tag{8}$$

On intègre les deux premières équations de (8) par rapport à τ sur $[0, 1]$; donc

$$\begin{aligned} &\frac{dx}{dt} = \int_0^1 g(\tau, x(t), u(t, \tau))\, d\tau, \\ &-\frac{dp}{dt} = \int_0^1 H_x(\tau, x(t), u(t, \tau), p(t))\, d\tau, \\ &H_v(\tau, x(t), u(t, \tau), p(t)) = 0. \end{aligned} \tag{9}$$

Mais on va montrer plus tard que les relations (9) ne sont rien d'autre que les conditions nécessaires d'optimalité du problème (6), (7). On dit que le problème (6), (7) est une sorte de généralisation des problèmes de contrôle optimal, car si l'on fait $\partial g/\partial \tau(\tau, x, v) \equiv 0$, $\partial l/\partial \tau(\tau, x, v) \equiv 0$ (un cas particulier de la fonction périodique), il devient un problème de contrôle optimal classique. On va voir que les méthodes de perturbations singulières jouent un rôle très important dans cet article. Ceci veut dire qu'on va souvent utiliser la technique de changement de l'échelle du temps et les techniques de développements asymptotiques. Pour démontrer le résultat de convergence, l'auteur a beaucoup profité des techniques utilisées initialement par Bensoussan [1], [2]. Le résultat de régularité d'ordre élevé de la fonction du coût (voir Peng [8]) est aussi nécessaire pour traiter la convergence d'ordre élevé de l'équation de Bellman avec vibrations rapides.

Dans § 2, on va traiter le principe du maximum pour le problème homogénéisé. Dans § 3 on va étudier la programmation dynamique du problème homogénéisé. Dans §§ 4 et 5, on va traiter, au point de vue du principe du maximum et de la programmation dynamique, respectivement, le problème de convergence d'ordre élevé en contrôle optimal avec vibrations rapides. On donnera un exemple simple dans § 6.

2. Principe de Pontryagin du problème homogénéisé. On suppose que

(10) g, l, h sont continûment différentiables par rapport à x, v,

$$\begin{aligned} &|g(\tau, x, v)| \leqq C(1+|x|+|v|), \qquad \left|\frac{\partial g}{\partial x}\right|, \left|\frac{\partial g}{\partial v}\right| \leqq C, \\ &|l(\tau, x, v)| \leqq C(1+|x|^2+|v|^2), \quad \left|\frac{\partial l}{\partial x}\right|, \left|\frac{\partial l}{\partial v}\right| \leqq C(1+|x|+|v|), \\ &\left|\frac{\partial h}{\partial x}(x)\right| \leqq C(1+|x|). \end{aligned} \tag{11}$$

Un contrôle admissible du problème homogénéisé est une fonction

(12) $v(\cdot,\cdot)\in L^2([0,T]\times[0,1];\mathbb{R}^k)$ t.q. $v(t,\tau)\in \mathbf{U}_{ad}$ p.p. dans $[0,T]\times[0,1]$

où

$$\mathbf{U}_{ad}\subset\mathbb{R}^k, \quad \text{convexe non vide.}$$

Dans ce chapitre on note que

$$EF(\cdot,x,v(t,\cdot))=\int_0^1 F(\tau,x,v(t,\tau))\,d\tau. \tag{13}$$

Le problème de contrôle optimal homogénéisé est

$$\frac{dx}{dt}=Eg(\cdot,x(t),v(t,\cdot)), \qquad \min_{v(\cdot,\cdot)}\left[E\int_0^T l(\cdot,x(t),v(t,\cdot))\,dt+h(x(T))\right]. \tag{14}$$

On a le principe de Pontryagin (voir le théorème 1).

THEOREME 1. *On fait les hypothèses* (10), (11). *Alors une condition nécessaire pour* $u(t,\tau)$ *soit un contrôle optimal du problème* (14) *est qu'il existe une fonction* $p(\cdot)\in H^1(0,T;\mathbb{R}^n)$, *telle que*

$$\begin{aligned}&\frac{dx}{dt}=Eg(\cdot,x(t),u(t,\cdot)), \qquad -\frac{dp}{dt}=EH_x(\cdot,x(t),u(t,\cdot),p(t)),\\ &H_v(\tau,x(t),u(t,\tau),p(t))(v-u(t,\tau))\geqq 0 \quad \forall t, \text{ p.p. } \forall v\in\mathbf{U}_{ad}.\end{aligned} \tag{15}$$

La démonstration est similaire à la méthode classique (voir, par exemple, Bensoussan [1]).

On aura besoin aussi des conditions du second ordre pour que $u(t,\tau)$ soit un contrôle optimal. On suppose en plus de (11), (12) que

$$g,\ l,\ h\in C^2(\mathbb{R}^n\times\mathbb{R}^k) \tag{16}$$

toutes les dérivées de g et les dérivées secondes de l, h sont supposées bornées.

On a le Théorème 2.

THEOREME 2. *On suppose* (11), (12), (16). *Soit* $(x(t),u(t,\tau),p(t))$, *un triplet vérifiant* (15) *et*

$$H_{vv}(\tau,x,v,p(t))\geqq\beta I, \quad \beta>0 \quad \forall x,\tau,t, \quad \forall v\in\mathbf{U}_{ad}, \tag{17}$$

$$(H_{xx}-H_{xv}H_{vv}^{-1}H_{vx})(\tau,x,v,p_0(t))\geqq 0 \tag{18}$$

pour les mêmes arguments qu'en (17); *alors* $u(t,\tau)$ *est l'unique contrôle optimal du problème* (14).

La démonstration est de nouveau similaire à la méthode classique (voir, par exemple, Bensoussan [1]).

3. Programmation dynamique du problème homogénéisé. On commence par la définition de la fonction de Bellman du problème homogénéisé. On considère pour $t_0\in[0,T]$, $x_0\in\mathbb{R}^n$

$$\frac{dx}{dt}=Eg(\cdot,x(t),v(t,\cdot)), \qquad x(t_0)=x_0. \tag{19}$$

La fonctionnelle à minimiser est donc paramétrée par la donnée initiale (x_0,t_0). La fonction de Bellman est définie par

$$\Phi(x_0,t_0)=\min_{v(\cdot,\cdot)} J^0_{x_0,t_0}(v(\cdot,\cdot)), \tag{20}$$

où

$$J^0_{x_0,t_0}(v(\cdot,\cdot)) = \int_0^T El(\cdot, x(s), v(s,\cdot))\, ds + h(x(T)).$$

On suppose

$$E|g(\cdot, x, v)|,\ E|l(\cdot, x, v)| < +\infty \quad \forall x \in \mathbb{R}^n, \quad v \in \mathbb{R}^k, \tag{21}$$

g, l, h sont continûment différentiable par rapport à x,

$$\mathbf{U}_{\text{ad}} \subset \mathbb{R}^k \text{ mesurable}, \tag{22a}$$

$$|g_x(\tau, x, v)| \leqq C, \tag{22b}$$

$$|h_x(x)| \leqq \bar{h} \cdot (1+|x|), \tag{22c}$$

$$|l_x(\tau, x, v)| \leqq \bar{l}(|l(\tau, 0, v)|^{1/2} + |x| + 1), \tag{22d}$$

$$l(\tau, x, v) \geqq l_0 |g(\tau, x, v) - g(\tau, x, \bar{v})|^2 - C_0, \tag{22e}$$

où

$$\bar{v} \in \mathbb{R}^k \text{ est fixé},$$

$$h(x) \geqq -C_0. \tag{23}$$

Remarque. Quand $\partial g/\partial \tau \equiv 0$, $\partial l/\partial \tau \equiv 0$, on revient au cas classique (voir [4], [1]). Même dans ce cas, les hypothèses (21)-(23) sont plus faibles que ce qu'on suppose souvent: on n'a plus besoin de la continuité par rapport à v, la condition de "quasilinéaire quadratique" de l par rapport à v est remplacée par (22e).

Le lemme suivant donne la régularité de la fonction de Bellman $\Phi(x, t)$.

LEMME 3. *Sous les hypothèses* (21)-(23), *la fonction* $\Phi(x, t)$ *vérifie*

$$|\Phi(x, t)| \leqq C(1+|x|^2), \tag{24}$$

$$|D\Phi(x, t)| \leqq \bar{C}(1+|x|), \tag{25}$$

$$\left|\frac{\partial \Phi}{\partial t}(x, t)\right| \leqq \bar{C}(1+|x|^2). \tag{26}$$

Démonstration. On prend

$$v(s) = \bar{u}, \quad t \leqq s \leqq T, \quad \bar{u} \in \mathbf{U}_{\text{ad}} \quad \text{fixé}$$

comme contrôle et on note $\bar{x}(s)$, la fonction d'état correspondante.

$$\frac{d\bar{x}(s)}{ds} = Eg(\cdot, \bar{x}(s), \bar{u}), \qquad \bar{x}(t) = x. \tag{27}$$

D'après (11) on a

$$|\bar{x}(s)| \leqq C(1+|x|) \quad \forall t \leqq s \leqq T. \tag{28}$$

Donc

$$\begin{aligned}
\Phi(x, t) &\leqq J_{x,t}(\bar{u}) \\
&= E \int_t^T l(\cdot, \bar{x}(s), \bar{u})\, ds + h(\bar{x}(T)) \\
&= E \int_t^T \left[l(\cdot, 0, \bar{u}) + \int_0^1 l_x(\cdot, \lambda \bar{x}(s), \bar{u}) \cdot \bar{x}(s)\, d\lambda \right] ds + h(\bar{x}(T)) \\
&\leqq \int_t^T (E|l(\cdot, 0, \bar{u})|\, ds + E \int_0^1 \bar{l}(|l(\cdot, 0, \bar{u})|^{1/2} + \lambda |\bar{x}(s)| + 1)\, d\lambda \cdot |\bar{x}(s)|)\, ds \\
&\quad + h(\bar{x}(T)) \quad \text{d'après (22d).}
\end{aligned}$$

De (28), (22), on a

$$\Phi(x, t) \leqq \bar{C}(1+|x|^2).$$

Mais d'après (22e) on sait que

$$\Phi(x, t) \geqq -C_0 T.$$

Les deux dernières inégalités impliquent donc l'estimation (24). Cela nous permet de limiter le contrôle admissible par

$$E \int_t^T l(\cdot, x(s), v(s, \cdot))\, ds \leqq C(1+|x|^2). \tag{29}$$

Ce qui avec l'hypothèse (22e) implique

$$l_0 \int_t^r E|g(\cdot, x(s), v(s, \cdot)) - g(\cdot, x(s), \bar{v})|^2\, ds \leqq C(1+|x|^2) \tag{30}$$

donc

$$|y(r)| \leqq C(1+|x|) \tag{31}$$

où on note que

$$y(r) = E \int_t^r (g(\cdot, x(s), v(s, \cdot)) - g(\cdot, x(s), \bar{v}))\, ds.$$

Mais on sait que

$$x(r) - x = y(r) + E \int_t^r g(\cdot, x(s), \bar{v})\, ds;$$

donc

$$|x(r)| \leqq C_1(1+|x|) + C_2 \int_t^r |x(s)|\, ds.$$

Par conséquent

$$|x(r)| \leqq C(1+|x|). \tag{32}$$

Ce qui avec (30) implique

$$E \int_t^T |g(\cdot, x(s), v(\cdot, s))|^2\, ds \leqq C(1+|x|^2). \tag{33}$$

On a aussi

$$E \int_t^T |g(\cdot, 0, v(s, \cdot))|^2\, ds \leqq C(1+|x|^2). \tag{34}$$

On peut montrer également que

$$\int_t^T |El(\cdot, x(s), v(\cdot, s))|\, ds \leqq C(1+|x|^2). \tag{35}$$

$$\int_t^T |El(\cdot, 0, v(\cdot, s))|\, ds \leqq C(1+|x|^2). \tag{36}$$

En effet, si on note

$$\Delta^+ = \{s \in [t, T], El(\cdot, x(s), v(s, \cdot)) > 0\}, \qquad \Delta^- = [t, T] \backslash \Delta^-.$$

On a d'après (29)

$$\int_{\Delta^+} |El(\cdot, x(s), v(s, \cdot))|\, ds - \int_{\Delta^-} |El(\cdot, x(s), v(s, \cdot))|\, ds \leqq C(1+|x|^2). \tag{36}_1$$

Toujours grâce à (22e) on a

$$|El(\cdot, x(s), v(\cdot, s))| \leqq C_0 \quad \forall s \in \Delta^-,$$

ce qui avec $(36)_1$ implique donc (35). On peut déduire (36) de (35) et l'hypothèse (22d).

Avec (32), (36), on peut montrer l'estimation (25). Pour une donnée initiale $\tilde{x} \in \mathbb{R}^n$, on note $\tilde{x}(t)$ la solution de

$$\frac{d\tilde{x}}{ds} = Eg(\cdot, \tilde{x}(s), v(s, \cdot)), \qquad \tilde{x}(t) = \tilde{x}.$$

On peut donc limiter $x(s)$, $\tilde{x}(s)$, $v(s, \tau)$, par

$$(36)_2 \qquad |x(s)|^2, |\tilde{x}(s)|^2, \int_t^T |El(\cdot, 0, v(s, \cdot))|\, ds \leqq C(1+|x|^2+|\tilde{x}|^2).$$

Par ailleurs on a

$$|x(s)-\tilde{x}(s)| \leqq |x-\tilde{x}| + E\int_t^s |g(\cdot, x(r), v(r, \cdot)) - g(\cdot, \tilde{x}(r), v(r, \cdot))|\, dr$$

$$\leqq |x-\tilde{x}| + C\int_t^s |x(r)-\tilde{x}(r)|\, dr.$$

D'après l'inégalité de Gronwall

$$(37) \qquad |x(s)-\tilde{x}(s)| \leqq C_1|x-\tilde{x}|.$$

On a donc

$$|J^0_{x,t}(v(\cdot)) - J^0_{\tilde{x},t}(v(\cdot))|$$

$$= \left| \int_t^T \int_0^1 El_x(\tau, \tilde{x}(s) + \lambda(x(s)-\tilde{x}(s)), v(s, \cdot))\, d\lambda\, (x(s)-\tilde{x}(s))\, ds \right|$$

$$+ \left| \int_0^1 h_x(\tilde{x}(T) + \lambda(x(T)-\tilde{x}(T)))\, d\lambda\, (x(T)-\tilde{x}(T)) \right|.$$

De (37) et les hypothèses (22c), (22d), cela est majoré par $\int_0^1 \int_t^T \bar{l}(|El(\cdot, 0, v(s, \cdot))|^{1/2} + |\tilde{x}(s) + \lambda(x(s)-\tilde{x}(s))| + 1)\, d\lambda\, ds. \quad |x-\tilde{x}| + \bar{h} \cdot |\tilde{x}(T) + \lambda(x(T)-\tilde{x}(T))| \cdot |x(T)-\tilde{x}(T)|$; de $(36)_2$, (37) on a finalement

$$|J^0_{x,t}(v(\cdot)) - J^0_{\tilde{x},t}(v(\cdot))| \leqq C|x-\tilde{x}| \cdot (1+|x|+|\tilde{x}|).$$

On en déduit donc

$$(38) \qquad |\Phi(x, t) - \Phi(\tilde{x}, t)| \leqq C|x-\tilde{x}|(1+|x|+|\tilde{x}|).$$

Cela implique que Φ est presque partout différentiable et que la dérivée $D\Phi(x, t)$ est majorée par (25).

L'estimation (26) est une conséquence du principe d'optimalité de la programmation dynamique homogénéisée.

$$(39) \qquad \Phi(x, t) = \inf_{v(\cdot,\cdot)} \left[\int_t^{t+h} El(\cdot, x(s), v(s, \cdot))\, ds + \Phi(x(t+h), t+h) \right].$$

Si l'on choisit $v(s, \tau) = \bar{u}$, $t \leqq s \leqq t+h$, et note $\bar{x}(t)$ la fonction d'état correspondante, on a d'après (39)

$$\Phi(x, t) \leqq \int_t^{t+h} El(\cdot, \bar{x}(s), \bar{u})\, ds + \Phi(\bar{x}(t+h), t+h). \tag{40}$$

De (28)

$$|\bar{x}(t+h) - x| \leqq C|h|(1+|x|).$$

Compte tenu de l'estimation pour $D\Phi$, on a

$$|\Phi(\bar{x}(t+h), t+h) - \Phi(x, t+h)| \leqq C|h|(1+|x|^2),$$

ce qui avec (40) implique

$$\Phi(x, t) \leqq C|h|(1+|x|^2) + \Phi(x, t+h)$$

donc

$$\frac{\partial \Phi}{\partial t} \geqq -C(1+|x|^2). \tag{41}$$

Par ailleurs on a

$$\begin{aligned} |\Phi(x(t+h), t+h) - \Phi(x, t+h)| &\leqq C|x(t+h) - x|(1+|x|) \\ &= C(1+|x|)|\cdot \left| \int_t^{t+h} Eg(\cdot, x(s), v(s, \cdot))\, ds \right|. \end{aligned}$$

On déduit alors de (39) tenant compte de l'hypothèse (22e)

$$\begin{aligned} \Phi(x, t) &\geqq \inf_{v(\cdot,\cdot)} \left[\int_t^{t+h} E|g(\cdot, x(s), v(s, \cdot)) - g(\cdot, x(s), \bar{v})|^2 \right] ds \\ &\quad - C_0|h| - C(1+|x|) \int_t^{t+h} E|g(\cdot, x(s), v(s, \cdot))|\, ds + \Phi(x, t+h) \\ &\geqq \inf_{v(\cdot,\cdot)} \left[\int_t^{t+h} E(|g(\cdot, x(s), v(s, \cdot))|^2 - C(1+|x|)|g(\cdot, x(s), v(s, \cdot))|) \right] ds \\ &\quad - C_0|h| + \Phi(x, t+h) \\ &\geqq \Phi(x, t+h) - C_1(1+|x|^2)|h| \end{aligned}$$

ce qui avec (41) implique enfin (26). □

L'équation de la programmation dynamique homogénéisée s'écrit

$$\frac{\partial \Phi}{\partial t} + \inf_{u(\cdot) \in \mathbf{U}_{ad}} E[D\Phi \cdot g(\cdot, x, v(\cdot)) + l(\cdot, x, v(\cdot))] = 0, \qquad \Phi(x, T) = h(x) \tag{42}$$

où

$$\mathbf{U}_{ad} = \{v(\cdot), v(\tau) \in L^2([0, T]; \mathbf{U}_{ad})\}. \tag{43}$$

Remarque. L'équation (42) est aussi une équation aux dérivées partielles du premier ordre. Elle peut s'écrire comme

$$\frac{\partial \Phi}{\partial t} + E[D\Phi g(\cdot, x, u(\cdot, x, D\Phi)) + l(\cdot, x, u(\cdot, x, D\Phi))] = 0, \qquad \Phi(x, T) = h(x) \tag{43}_1$$

où

$$(43)_2 \qquad u(\tau, x, D\Phi) = \arg \inf_{u(\cdot)\in U_{ad}} E[D\Phi \cdot g(\cdot, x, u(\cdot)) + l(\cdot, x, u(\cdot))],$$

la fonction $\hat{u}(\tau, x) = u(\tau, x, D\Phi(x))$, s'appelle le contrôle feedback. Si $\hat{u}(\tau, x)$ est Lipschitzien par rapport à x, on peut montrer que $\hat{u}(\tau, x(t))$, $0 \leqq t \leqq \tau$, est exactement le contrôle optimal du problème (6), (7), où $x(t)$ est l'état correspondant

$$(44) \qquad \frac{dx}{dt} = Eg(\cdot, x(t), v(\cdot, x(t))), \qquad x(0) = x_0.$$

Maintenant on peut énoncer le théorème 4.

THEOREME 4. *Sous les hypothèses* (21)-(23). *La fonction* $\Phi(x, t)$ *définie par* (16) *vérifie* (42). *C'est de plus la fonction maximum vérifiant* (42) *et le lemme* 3.

Démonstration. On peut se limiter aux contrôles vérifiant

$$(45) \qquad \int_t^{t+h} E|g(\cdot, 0, v(s, \cdot))|^2 \, ds \leqq Ch(1+|x|^2).$$

En effet on peut considérer seulement les contrôles vérifiant

$$(46) \qquad \int_t^{t+h} El(\cdot, x(s), v(s, \cdot)) \, ds + \Phi(x(t+h), t+h)$$

$$\leqq \int_t^{t+h} El(\cdot, \bar{x}(s), \bar{u}) \, ds + \Phi(\bar{x}(t+h), t+h)$$

où $(\bar{x}(s), \bar{u})$ est le même couple dans (40). D'après le lemme 3 on a

$$|\Phi(x(t+h), t+h) - \Phi(x, t)| \leqq C|x(t+h) - x|(1+|x|) + C(1+|x|^2)h,$$

$$|\Phi(\bar{x}(t+h), t+h) - \Phi(x, t)| \leqq C(1+|x|^2)h.$$

Ce qui avec (46) implique donc

$$\int_t^{t+h} El(\cdot, x(s), v(s, \cdot)) \, ds$$

$$\leqq C|x(t+h) - x|(1+|x|) + C(1+|x|^2)h$$

$$\leqq C_1(1+|x|^2)h + C(1+|x|) \int_t^{t+h} E|g(\cdot, 0, v(s, \cdot))| \, ds$$

$$\leqq C_1(1+|x|^2)h + C(1+|x|)h^{1/2}\left(\int_t^{t+h} E|g(\cdot, 0, v(s, \cdot))|^2 \, ds\right)^{1/2}.$$

D'autre part d'après (22e) on a

$$\int_t^{t+h} l_0 E|g(\cdot, x(s), v(s, \cdot)) - g(\cdot, x(s), \bar{v})|^2 \, ds$$

$$\leqq C_2(1+|x|^2)h + C(1+|x|)h^{1/2}\left(\int_t^{t+h} E|g(\cdot, 0, v(s, \cdot))|^2 \, ds\right)^{1/2}.$$

D'après (32) et l'hypothèse (22b)

$$\int_t^{t+h} l_0 E|g(\cdot, 0, v(s,\cdot))|^2\, ds \leqq C_3(1+|x|^2)h + C(1+|x|)h^{1/2}\left(\int_t^{t+h} E|g(\cdot, 0, v(s,\cdot))|^2\, ds\right)^{1/2};$$

cela implique donc (45).

Pour les contrôles vérifiant (45), on a

$$|x(t+h)-x| \leqq Ch(1+|x|) \tag{47}$$

où C ne dépend pas du contrôle dans la classe (45).

Soit maintenant x, t est un point où Φ est différentiable, on a

$$\left|\Phi(x(t+h), t+h)-\Phi(x,t)-h\frac{\partial\Phi}{\partial t}-\int_t^{t+h} ED\Phi\cdot g(\cdot, x, v(s,\cdot))\, ds\right| \leqq O(h)\cdot h$$

où $O(h)$ tend vers zéro avec h et ne dépend pas du contrôle. Par conséquent de (16) il résulte

$$\inf_{v(\cdot,\cdot)}\left[\frac{1}{h}\int_t^{t+h} E(D\Phi\cdot g(\cdot, x, v(s,\cdot))+l(\cdot, x, v(s,\cdot)))\, ds\right]+\frac{\partial\Phi}{\partial t}\leqq O(h)$$

donc

$$\inf_{v\in \mathbf{U}_{ad}} E[D\Phi g(\cdot, x, v(\cdot))+l(\cdot, x, v(\cdot))]+\frac{\partial\Phi}{\partial t}\leqq O(h).$$

En faisant $h\to 0$

$$\inf_{v\in \mathbf{U}_{ad}} \mathrm{E}[\mathrm{D}\Phi \mathrm{g}(\cdot, \mathrm{x}, \mathrm{v}(\cdot))+\mathrm{l}(\cdot, \mathrm{x}, \mathrm{v}(\cdot))]+\frac{\partial\Phi}{\partial \mathrm{t}}\leqq 0. \tag{48}$$

Comme par ailleurs, toujours d'après (39) on a

$$\Phi(x,t)\leqq \int_t^{t+h} El(\cdot, x(s), v(\cdot))\, ds+\Phi(x(t+h), t+h)$$

$$= El(\cdot, x, v(\cdot))\cdot h+\frac{\partial\Phi}{\partial t}\cdot h+D\Phi Eg(\cdot, x, v(\cdot))h+\Phi(x,t)+O(h).$$

On a

$$\frac{\partial\Phi}{\partial t}+E[D\Phi g(\cdot, x, v(\cdot))+l(\cdot, x, v(\cdot))]\geqq 0$$

d'où

$$\inf_{v(\cdot)\in \mathbf{U}_{ad}} E[D\Phi g(\cdot, x, v(\cdot))+l(\cdot, x, v(\cdot))]+\frac{\partial\Phi}{\partial t}\geqq 0$$

ce qui avec (48) implique (42).

On peut montrer l'affirmation "solution maximum" de façon similaire à Bensoussan [1]. □

On peut traiter la régularité de $\Phi(x, t)$ pour les ordres plus élevés de la même façon qu'en Peng [8, Chap. I] (l'étude de la régularité de l'équation de Bellman par la méthode de perturbation). On énonce ici seulement le résultat.

THEOREME 5. *On fait les hypothèses*

(49) (i) g, l, h *sont* $2p+2$ *fois continûment différentiables*;

(ii) $$\left|\frac{\partial^{(j)} g}{(\partial\beta)^j}(\tau, \beta)\right| \leqq C_j(1+|x|)^{1-j},$$

$$\left|\frac{\partial^{(j)} l}{(\partial\beta)^j}(\tau, \beta)\right|, \left|\frac{\partial^{(j)} h}{(\partial x)^j}(x)\right| \leqq C(1+|x|)^{2-j} \quad \forall 0 \leqq j < 2p+2$$

où on note

$$\beta = (x, v),$$

(iii) $$\frac{\partial^2 h}{\partial x^2} \geqq 0, \qquad h(x) \geqq -C,$$

(iv) $$\frac{\partial^2 l}{\partial \beta^2}(\tau, \beta) \geqq \gamma_0 I$$

où γ_0 *est assez grand par rapport à* C_2,

(v) $\mathbf{U}_{\text{ad}} = \mathbb{R}k$.

On a alors

(50) (i) $\Phi(x, t) \in C^{2p+1}$ *et*

$$\left|D_x^{(j)} \frac{\partial^{(i)}}{\partial t^i} \Phi(x, t)\right| \leqq C(1+|x|)^{2-j}, \qquad i+j \leqq 2p+1,$$

(ii) *le contrôle feedback optimal* $u(x, t, \tau)$ *existe; il est caractérisé par*

$$D\Phi g_v(\tau, x, u(x, t, \tau)) + l_v(\tau, x, u(x, t; \tau)) = 0,$$

(iii) $$\left|\frac{\partial^{(i+j)}}{\partial t^i \partial x^j} u(x, t, \tau)\right| \leqq C(1+|x|)^{1-j}, \qquad i+j \leqq 2p.$$

4. Méthode asymptotique du point de vue du principe maximum.

4.1. Développement asymptotique. On va étudier dans ce paragraphe la méthode des perturbations pour le système d'optimalité, du point de vue du principe de Pontryagin. On va prendre le cas sans contrainte; $\mathbf{U}_{\text{ad}} = \mathbb{R}^K$. Dans ce cas, les conditions nécessaires pour le problème (3) seront

$$\frac{dx^\varepsilon}{dt} = g\left(\frac{t}{\varepsilon}, x^\varepsilon, u^\varepsilon\right), \qquad x^\varepsilon(0) = x_0,$$

(51)
$$-\frac{dp^\varepsilon}{dt} = H_x\left(\frac{t}{\varepsilon}, x^\varepsilon, u^\varepsilon\right), \qquad p^\varepsilon(T) = h(x^\varepsilon(T)),$$

$$H_v\left(\frac{t}{\varepsilon}, x^\varepsilon, u^\varepsilon\right) = 0.$$

On va voir que, quand $\varepsilon \to 0$, ce système tend vers le système d'homogénéisation (6), (7). Pour les développements d'ordre élevé, on utilise naturellement la méthode de changement de l'échelle du temps, $\tau = t/\varepsilon$. Mais on va rencontrer à ce niveau là un

problème. Premièrement, à la différence des perturbations singulières, les vibrations rapides ne disparaissent pas avec le déroulement du temps. De plus, il est évident que $\tau = t/\varepsilon$ ne vérifie pas la condition au bord $t = T$, tandis que $\tau = (T-t)/\varepsilon$ ne la vérifie pas pour $t = 0$. Pour résoudre cette difficulté, on va essayer de développer x^ε, p_ε, u^ε sous la forme suivante

$$
\begin{aligned}
x^\varepsilon(t) &= x_0(t) + \varepsilon x_1\left(t, \frac{t}{\varepsilon}\right) + \cdots, \\
p^\varepsilon(t) &= p_0(t) + \varepsilon p_1\left(t, \frac{t}{\varepsilon}\right) + \cdots, \qquad (52) \\
u^\varepsilon(t) &= u_0\left(t, \frac{t}{\varepsilon}\right) + \varepsilon u_1\left(t, \frac{t}{\varepsilon}\right) + \cdots.
\end{aligned}
$$

On considère $\alpha_0 = T/\varepsilon$ comme un paramètre. En remplaçant x^ε, p^ε, u^ε par (52) dans (51) et en général les termes du même ordre, on a

$$
\begin{aligned}
&\frac{dx_0}{dt} + \frac{dx_1}{d\tau} = g(\tau, x_0(t), u_0(t, \tau)), \\
&\frac{dp_0}{dt} - \frac{\partial x_1}{\partial \tau} = H_x(\tau, x_0(t), u_0(t, \tau), p_0(t)), \qquad (53) \\
&H_v(\tau, x_0(t), u_0(t, \tau), p_0(t)) = 0
\end{aligned}
$$

avec les conditions au bord

$$
x_0(0) = x_0, \qquad p_0(T) = h_x(x_0(T)), \qquad (54)
$$

$$
x_1(0, 0) = 0, \qquad p_1(T, \alpha_0) = h_{xx}(x_0(T))x_1(T, \alpha_0). \qquad (55)
$$

Pour les relations du deuxième ordre on a

$$
\begin{aligned}
&\frac{\partial x_1}{\partial t} + \frac{\partial x_2}{\partial \tau} = g_x(\tau, x_0(t), u_0(t, \tau))x_1 + g_v(\tau, x_0, u_0)u_1, \\
&-\frac{\partial p_1}{\partial t} - \frac{\partial p_2}{\partial \tau} = g_x^*(\cdot)p_1 + H_{xx}(\cdot)x_1 + H_{xv}(\cdot)u_1, \qquad (56) \\
&g_v^*(\cdot)p_1 + H_{vx}(\cdot)x_1 + H_{vv}(\cdot)u_1 = 0
\end{aligned}
$$

avec les conditions au bord

$$
x_2(0, 0) = 0, \qquad p_2(T, \alpha_0) = \tfrac{1}{2}h_x^{(3)}(x_0(T))(x_1(T, \alpha_0))^2 + h_{xx}(\cdot)x_2(T, \alpha_0). \qquad (57)
$$

Généralement, si on note

$$
\beta_i = (x_i, v_i), \qquad \sigma_i = (x_i, v_i, p_i)
$$

pour un nombre entier $l \geqq 1$, on a

$$
\begin{aligned}
&\frac{\partial x_l}{\partial t} + \frac{\partial x_{l+1}}{\partial \tau} = \sum_{(\alpha \cdot l)}^{l} \frac{1}{\alpha_1! \cdots \alpha_l!} g_\beta^{(\alpha_1 + \cdots + \alpha_l)} \beta_1^{\alpha_1} \cdots \beta_l^{\alpha_l}, \\
&-\frac{\partial p_l}{\partial t} - \frac{\partial p_{l+1}}{\partial \tau} = \sum_{(\alpha \cdot l)}^{l} \frac{1}{\alpha_1! \cdots \alpha_l!} D_\sigma^{(\alpha_1 + \cdots + \alpha_l)} H_x(\tau, \sigma_0) \sigma_1^{\alpha_1} \cdots \sigma_l^{\alpha_l}, \qquad (58) \\
&\sum_{(\alpha \cdot l)}^{l} \frac{1}{\alpha_1! \cdots \alpha_l!} D_\sigma^{(\alpha_1 + \cdots + \alpha_l)} H_v(\tau, \sigma_0) \sigma_1^{\alpha_1} \cdots \sigma_l^{\alpha_l} = 0.
\end{aligned}
$$

On peut aussi écrire (58) sous la forme suivante:

$$
\begin{aligned}
&\frac{\partial x_l}{\partial t}+\frac{\partial x_{l+1}}{\partial \tau}=g_x(\cdot)x_l+g_v(\cdot)u_l \\
&\qquad\qquad +\sum_{(\alpha,l-1)}^{l}\frac{1}{\alpha_1!\cdots\alpha_{l-1}!}g_\beta^{(\alpha_1+\cdots+\alpha_{l-1})}\beta_1^{\alpha_1}\cdots\beta_{l-1}^{\alpha_{l-1}}, \\
&-\frac{\partial p_l}{\partial t}-\frac{\partial p_{l+1}}{\partial \tau}=g_x^*(\cdot)p_l+H_{xx}(\cdot)x_l+H_{xv}(\cdot)u_l \\
&\qquad\qquad +\sum_{(\alpha,l-1)}^{l}\frac{1}{\alpha_1!\cdots\alpha_{l-1}!}D_\sigma^{(\alpha_1+\cdots+\alpha_{l-1})}H_x(\cdot)\sigma_1^{\alpha_1}\cdots\sigma_{l-1}^{\alpha_{l-1}}, \\
&0=g_v^{*(\cdot)}p_l+H_{vx}(\cdot)x_l+H_{vv}(\cdot)U_l \\
&\qquad\qquad +\sum_{(\alpha,l-1)r}^{l}\frac{1}{\alpha_1!\cdots\alpha_{l-1}!}D_\sigma^{(\alpha_1+\cdots+\alpha_{l-1})}H_x(\cdot)\sigma_1^{\alpha_1}\cdots\sigma_{l-1}^{\alpha_{l-1}}
\end{aligned}
\tag{59}
$$

avec les conditions au bord

$$
\begin{aligned}
&x_l(0,0)=0, \\
&p_l(T,\alpha_0)=\sum_{(\alpha\cdot l)}^{l}\frac{1}{\alpha_1!\cdots\alpha_l!}D_x^{(1+\alpha_1+\cdots+\alpha_l)}h(x_0(T))x_1^{\alpha_1}(T,\alpha_0)\cdots xl^{\alpha_l}(T,\alpha_0)
\end{aligned}
\tag{60}
$$

où

$$
\sum_{(\alpha,l)}^{j} \text{ désigne la somme } \sum_{\substack{\alpha_1+2\alpha_2+\cdots+l\alpha_l=j,\\ \alpha_1,\cdots,\alpha_l>0}}.
\tag{61}
$$

Remarque 6. Soit $F(\tau)$ une fonction de 1-périodique. On suppose que $\int_0^1|F(\tau)|\,d\tau\leqq C$. Alors, pour que l'équation suivante ait une solution

$$
\frac{dy(\tau)}{d\tau}=F(\tau),\quad y(\tau)\text{: 1-périodique},\quad y(\tau_0)=y_0\in\mathbb{R}^n,
\tag{62}
$$

la condition nécessaire et suffisante est

$$
\int_0^1 F(\tau)\,d\tau=0.
\tag{63}
$$

De plus sous (63), la solution de (62) est unique. □

4.2. Résolution et estimation des termes du développement. On va résoudre progressivement x_i, u_i, p_i, $i=0,1,\cdots$ par (53), (56), (58)-(60). On considère tout d'abord x_0, u_0, p_0.

D'après la remarque 6 pour que $x_1(t,\tau)$, $p_1(t,\tau,\alpha_0)$ soient périodiques par rapport à τ, on doit avoir nécessairement

$$
\begin{aligned}
&\frac{dx_0}{dt}=\int_0^1 g(\tau,x(t),u_0(t,\tau))\,d\tau, \\
&-\frac{dp_0}{dt}=\int_0^1 H_x(\tau,x_0(t),u_0(t,\tau),p_0(t))\,d\tau,\qquad H_v(\tau,x_0(t),u_0(t,\tau),p_0(t))=0
\end{aligned}
\tag{64}
$$

avec les conditions au bord

$$
x_0(0)=x_0,\qquad p_0(T)=h_x(x_0(T)).
\tag{65}
$$

D'après le théorème 1, on sait que (64), (65) sont des conditions nécessaires pour que $u_0(t, \tau)$ soit un contrôle optimal du problème (14) avec $\mathbf{U}_{ad} = \mathbb{R}^k$. Sous les hypothèses du théorème 2, $u_0(t, \tau)$ est l'unique contrôle optimal du problème. Pour l'obtention de l'existence et de la régularité des termes d'ordre plus élevé, on suppose que

(66) $g(\tau, x, v)$, $\mathfrak{L}(\tau, x, v)$, $h(x)$ sont $C^{K+2}(\mathbb{R}^n \times \mathbb{R}^k)$ par rapport à x, v.

On peut vérifier aisément le lemme 7.

LEMME 7. *On suppose les hypothèses du théorème 2 satisfaites et également* (66). *Alors* $(dx_0/dt)(t)$, $(dp_0/dt)(t)$, $u_0(t, \tau)$ *sont de classe* C^{K+1} *par rapport à* t.

Pour trouver x_1, p_1, on les décompose sous la forme

$$x_1(t, \tau) = Y_1(t, \tau) + y_1(t), \qquad p_1(t, \tau) = Q_1(t, \tau) + q_1(t) \tag{67}$$

où, ayant satisfait les conditions au bord (60), on pose

$$Y_1(t, 0) = 0, \qquad y_1(0) = 0, \tag{68}$$

$$Q_1(t, \alpha_0) = 0, \qquad q_1(T, \alpha_0) = h_{xx}(x_0(T))(y_1(T) + Y_1(T, \alpha_0)). \tag{69}$$

Maintenant, on peut résoudre (53) avec les décompositions (67) et les conditions au bord (68), (69). On a

$$Y_1(t, \tau) = -\frac{dx_0}{dt}\tau + \int_0^\tau g(s, \beta_0(t, s))\, ds, \tag{70}$$

$$Q_1(t, \tau, \alpha_0) = -\frac{dp_0}{dt}(\tau - \alpha_0) - \int_{\alpha_0}^\tau H_x(s, \sigma_0(t, s))\, ds. \tag{71}$$

De plus, du lemme 7, on déduit que

$$Y_1(t, \tau),\ Q_1(t, \tau),\ \frac{\partial Y_1}{\partial \tau},\ \frac{\partial Q_1}{\partial \tau} \quad \text{sont de classe } C^{K+1} \text{ par rapport à } t. \tag{72}$$

Maintenant on peut trouver $y_1(t)$, $q_1(t)$ et $u_1(t, \tau)$. En appliquant la remarque 6 à (56), on sait que nécessairement on a

$$\begin{aligned}
\frac{dy_1}{dt} &= \int_0^1 g_x(\tau, x_0(t), u_0(t, \tau))\, dt\, y_1(t) + \int_0^1 g_v(\tau, x_0, u_0) u_1(t, \tau)\, d\tau \\
&\quad + \int_0^1 \left[-\frac{\partial Y_1}{\partial t} g_x(\tau, x_0, u_0)\, Y_1(t, \tau) \right] d\tau, \\
&\qquad\qquad y_1(0) = 0, \\
-\frac{dq_1}{dt} &= \int_0^1 g_x^*(\tau, \beta_0(t, \tau))\, d\tau \cdot q_1(t) + \int_0^1 H_x(\tau, \sigma_0(t, \tau))\, d\tau \cdot y_1(t) \\
&\quad + \int_0^1 H_v(\cdot) u_1(t, \tau)\, d\tau \\
&\quad + \int_0^1 \left[\frac{\partial Q_1}{\partial t} + g_x^*(\cdot) Q_1(t, \tau) + H_{xx}(\cdot)\, Y_1(t, \tau) \right] d\tau, \\
q_1(T) &= h_{xx}(x_0(T))(y_1(T) + Y_1(T, \alpha_0)), \\
g_v^*(\tau, \beta_0(t, \tau)) q_1(t) &+ H_{vx}(\tau, \sigma_0(t, \tau)) y_1(t) \\
&\quad + H_{vv}(\cdot) u_1(t, \tau) + g_v^*(\cdot) Q(t, \tau) + H_{vx}(\cdot)\, Y_1(t, \tau) = 0
\end{aligned} \tag{73}$$

ou

$$\frac{dy_1}{dt} = \int_0^1 (A(t,\tau)y_1(t) + B(t,\tau)u_1(t,\tau) + C(t,\tau))\, d\tau,$$

$$-\frac{dq_1}{dt} = \int_0^1 (A^*(t,\tau)q_1(t) + H_{11}(t,\tau)y_1(t) + H_{12}(t,\tau)u_1(t,\tau) + D(t,\tau))\, d\tau, \tag{74}$$

$$B^*(t,\tau)q_1(t) + H_{21}(t,\tau)y_1(t) + H_{22}(t,\tau)u_1(t,\tau) + E(t,\tau) = 0,$$

$$y_1(0) = 0, \qquad q_1(T) = h_{xx}(x_0(T))(y_1(T) + Y_1(T,\alpha_0))$$

où l'on note

$$A(t,\tau) = g_x(\tau, \beta_0(t,\tau)), \qquad B(t,\tau) = g_v(\tau, \beta_0(t,\tau)),$$

$$\begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix}(t,\tau) = \begin{pmatrix} H_{xx} & H_{xv} \\ H_{vx} & H_{vv} \end{pmatrix}(t, \sigma_0(t,\tau)),$$

$$C(t) = -\frac{\partial Y_1}{\partial t} + A(t,\tau)Y_1(t,\tau),$$

$$D(t) = \frac{\partial Q_1}{\partial t} + A^*(t,\tau)Q_1(t,\tau) + H_{11}(t,\tau)Y_1(t,\tau),$$

$$E(t,\tau) = B^*(t,\tau)Q_1(t,\tau) + H_{21}(t,\tau)Y_1(t,\tau).$$

On peut relier (73) à un problème linéaire quadratique

$$\frac{dy_1}{dt} = \int_0^1 (A(t,\tau)y_1(t) + B(t,\tau)v(t,\tau))\, d\tau$$

$$\min \left[\frac{1}{2} \int_0^T \int_0^1 \begin{pmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{pmatrix} \begin{pmatrix} y_1(t) \\ v_1(t,\tau) \end{pmatrix} + D(t,\tau)y_1(t) \right. \tag{75}$$

$$\left. + (E(t,\tau)v_1(t,\tau))\, d\tau\, dt + \frac{1}{2} h_{xx}(x_0(t))(y_1(T) + Y_1(T,\alpha_0))^2 \right].$$

De la même façon, on peut résoudre $x_2(t,\tau,\alpha_0)$, $p_2(t,\tau,\alpha_0)$, $u_2(t,\tau,\alpha_0)\cdots$, $x_K(t,\tau,\alpha_0)$, $p_K(t,\tau,\alpha_0)$, $u_K(t,\tau,\alpha_0)$ successivement et on a le lemme 8.

LEMME 8. *Sous les mêmes hypothèses que le lemme 7, on peut résoudre* x_i, u_i, p_i, $i = 0, 1, \cdots, K$ *solution de* (59). *De plus on a*

$$\frac{\partial x_i}{\partial t}, \frac{\partial p_i}{\partial t}, \frac{\partial^2 x_i}{\partial \tau\, \partial t}, \frac{\partial^2 p_i}{\partial \tau\, \partial t}, u_i(t,\tau) \quad \textit{sont de classe } C^{K+1-i} \tag{76}$$

par rapport à t.

On peut également définir $Y_{K+1}(t,\tau)$, $Q_{K+1P}(t,\tau)$ par

$$Y_{K+1}(t,\tau) = \int_0^\tau \left(-\frac{\partial x_K}{dt} + \sum_{(\alpha,K)}^{K} \frac{1}{\alpha_1! \cdots \alpha_K!} D_\beta^{\alpha_1+\cdots\alpha_K} g(s,\beta_0)\beta_1^{\alpha_1} \cdots \beta_K^{\alpha_K} \right) ds,$$

$$Q_{K+1}(t,\tau) = \int_0^\tau \left(-\frac{\partial p_K}{dt} + \sum_{(\alpha,K)}^{K} \frac{1}{\alpha_1! \cdots \alpha_K!} D_\sigma^{\alpha_1+\cdots\alpha_K} H_x(s,\sigma_0)\sigma_1^{\alpha_1} \cdots \sigma_K^{\alpha_K} \right) ds,$$

et

$$Y_{K+1}, Q_{K+1}, \frac{\partial Y_{K+1}}{d\tau}, \frac{\partial Q_{K+1}}{d\tau} \quad \text{sont } C^1 \text{ par rapport à } t. \tag{76'}$$

4.3. Résultats de convergence. Maintenant on va se demander si on peut approcher le problème périodique par la solution du développement asymptotique. On utilisera la méthode de Bensoussan [1].

On note

$$\bar{x}(t)=\sum_{l=0}^{K} x_l\left(t,\frac{t}{\varepsilon}\right)\varepsilon^l, \qquad \bar{u}(t)=\sum_{l=0}^{K} u_l\left(t,\frac{t}{\varepsilon}\right)\varepsilon^l,$$

$$\bar{p}(t)=\sum_{l=0}^{K} p_l\left(t,\frac{t}{\varepsilon}\right)\varepsilon^l, \qquad \hat{x}(t)=x^\varepsilon(t)-\bar{x}(t),$$

$$\hat{u}(t)=v(t)-\bar{u}(t), \qquad \bar{\beta}(t)=(\bar{x}(t),\bar{u}(t)),$$

$$\hat{\beta}(t)=(\hat{x}(t),\hat{u}(t)), \qquad \bar{\sigma}(t)=(\bar{x}(t),\bar{u}(t),\bar{p}(t))$$

où $v(t)$ est un contrôle admissible et $x^\varepsilon(t)$ est l'état correspondant.

$$\frac{dx^\varepsilon}{dt}=g\left(\frac{t}{\varepsilon},x^\varepsilon(t),v(t)\right), \qquad x^\varepsilon(0)=x_0. \tag{77}$$

On a le lemme 9.

LEMME 9. *On fait les hypothèses du théorème* 2 *et* (66). *Alors on a*

$$\begin{aligned} J^\varepsilon(v(\cdot))=&\int_0^T\left[H\left(\frac{t}{\varepsilon},\bar{\sigma}(t)\right)-\bar{p}(t)\bar{x}(t)\right]dt+h(\bar{x}(T))\\ &+\int_0^T\int_0^1\int_0^1 \lambda H_{\beta\beta}\left(\frac{s}{\varepsilon},\bar{\beta}(t)+\lambda\mu\hat{\beta}(t),\bar{p}(t)\right)\hat{\beta}^2(t)\,d\lambda\,d\mu\,dt\\ &+\varepsilon^{K+1}O(|\hat{x}(T)|)+\varepsilon^{K+1}O(|\hat{\beta}|_{L^2(O,T)}). \end{aligned} \tag{78}$$

Démonstration. On a

$$\begin{aligned} J^\varepsilon(v(\cdot))&=\int_0^T l\left(\frac{t}{\varepsilon},x^\varepsilon(t),v(t)\right)dt+h^\varepsilon(x^\varepsilon(T))\\ &=\int_0^T H\left(\frac{t}{\varepsilon},x^\varepsilon(t),v(t),\bar{p}(t)\right)dt-\int_0^T p(t)\frac{dx^\varepsilon}{dt}\,dt+h(x^\varepsilon(T))\\ &=\int_0^T\left[H\left(\frac{t}{\varepsilon},\bar{\sigma}(t)\right)+H_\beta\left(\frac{t}{\varepsilon},\bar{\sigma}(t)\right)\hat{\beta}(t)\right.\\ &\quad\left.+\int_0^1\int_0^1\lambda H_{\beta\beta}\left(\frac{t}{\varepsilon},\bar{\beta}(t)+\lambda\mu\hat{\beta}(t)\bar{p}(t)\hat{\beta}^2(t)\right)d\lambda\,d\mu\right]dt\\ &\quad-\int_0^T\bar{p}(t)\frac{dx^\varepsilon}{dt}\,dt+h(x^\varepsilon(T)) \end{aligned} \tag{79}$$

$$\begin{aligned} &=\int_0^T\left(H\left(\frac{t}{\varepsilon},\bar{\sigma}(t)\right)-\bar{p}(t)\frac{d\bar{x}}{dt}\right)dt+h(\bar{x}(T))\\ &\quad+\int_0^T\int_0^1\int_0^1\lambda H_{\beta\beta}(\cdot)\hat{\beta}^2(t)\,d\lambda\,d\mu\,dt\\ &\quad+\int_0^T H_\beta\left(\frac{t}{\varepsilon},\bar{\sigma}(t)\right)\hat{\beta}(t)\,dt-\int_0^T\bar{p}(t)\frac{d\hat{x}}{dt}\,dt\\ &\quad+h_x(\bar{x}(T))\hat{x}(T)+\int_0^1\int_0^1\lambda h_{xx}(\bar{x}(T)+\lambda\mu\hat{x}(T))\hat{x}^2(T)\,d\mu\,d\lambda. \end{aligned}$$

Pour la 3ème intégrale à droite on a

$$
\begin{aligned}
&[H_v(\tau, \bar{\sigma}(t))]_{\langle l\rangle} = 0, \qquad l = 0, 1, \cdots, K,\\
&[H_x(\tau, \bar{\sigma}(t))]_{\langle l\rangle} = -\frac{\partial p_l}{\partial t} - \frac{\partial p_{l+1}}{\partial \tau}, \qquad l = 0, 1, \cdots, K-1,\\
&[H_x(\tau, \bar{\sigma}(t))]_{\langle k\rangle} = -\frac{\partial p_k}{\partial t} - \frac{\partial Q_{K+1}}{\partial \tau},
\end{aligned} \tag{80}
$$

où l'on note

$$
\left[\sum_{i=1}^{K} b_i \varepsilon^i + O(\varepsilon^{K+1})\right]_{\langle l\rangle} = b_l, \qquad l = 1, \cdots, k. \tag{81}
$$

Donc

$$
\begin{aligned}
\int_0^T H_\beta\left(\frac{t}{\varepsilon}, \bar{\sigma}(t)\right)\hat{\beta}(t)\, dt &\geqq \int_0^T \left(-\frac{\bar{\partial} p}{dt} - \varepsilon^K \frac{\partial Q_{K+1}}{d\tau}\left(t, \frac{t}{\varepsilon}\right)\right)\hat{x}(t)\, dt\\
&\quad - C\varepsilon^{K+1}\int_0^T |\hat{\beta}(t)|\, dt\\
&= \int_0^T \bar{p}(t)\frac{d\hat{x}}{dt}\, dt - \bar{p}(T)\hat{x}(T) - C\varepsilon^{K+1}\int_0^T |\hat{\beta}(t)|\, dt\\
&\quad - \varepsilon^K \int_0^T \frac{\partial Q_{K+1}}{d\tau}\left(t, \frac{t}{\varepsilon}\right)\hat{x}(t)\, dt.
\end{aligned} \tag{82}
$$

Pour le dernier terme on a

$$
\frac{\partial Q_{K+1}}{\partial \tau}\left(t, \frac{t}{\varepsilon}\right) = \varepsilon\left(\frac{dQ_{K+1}}{dt}\left(t, \frac{t}{\varepsilon}\right) - \frac{\partial Q_{K+1}}{\partial t}\left(t, \frac{t}{\varepsilon}\right)\right);
$$

donc

$$
\begin{aligned}
&-\varepsilon^K \int_0^T \frac{\partial Q_{K+1}}{d\tau}\left(t, \frac{t}{\varepsilon}\right)\hat{x}(t)\, dt = \varepsilon^{K+1}\int_0^T Q_{K+1}(t, \tau)\frac{d\hat{x}}{dt}\, dt,\\
&\varepsilon^{K+1}\int_0^T \frac{\partial Q_{K+1}}{d\tau}\left(t, \frac{t}{\varepsilon}\right)\hat{x}(t)\, dt = \varepsilon^{K+1} O\left(\int_0^T |\hat{\beta}(t)|\, dt\right) \quad \text{(d'après (3.76)')}.
\end{aligned} \tag{83}
$$

De (79), (82), (83), tenant compte de l'hypothèse (18), on a

$$
\begin{aligned}
J^\varepsilon(v(\cdot)) &= \int_0^T \left(H\left(\frac{t}{\varepsilon}, \bar{\sigma}(t)\right) - \bar{p}(t)\frac{d\bar{x}}{dt}\right) dt + h(\bar{x}(T))\\
&\quad + \int_0^T \int_0^1 \int_0^1 \lambda H_{\beta\beta}(\cdot)\hat{\beta}^2(t)\, d\lambda\, d\mu\, dt + O(\varepsilon^{K+1})\int_0^T |\hat{\beta}(t)|\, dt\\
&\quad + h_x(\bar{x}(T))\cdot\hat{x}(T) - \bar{p}(T)\hat{x}(T).
\end{aligned} \tag{84}
$$

Mais on sait que

$$
[h_x(\bar{x}(T))]_{\langle l\rangle} = p_l(T, \alpha_0) \tag{85}
$$

ce qui avec (84) implique (78). □

On a aussi le lemme 10.

LEMME 10. *On a*

$$
|\bar{x}^\varepsilon(t) - \bar{x}(t)|_{C(0,T)} \leqq C\varepsilon^{K+1} \tag{86}
$$

où $x^\varepsilon(t)$ est la solution de

$$\frac{d\bar{x}^\varepsilon}{dt} = g\left(\frac{t}{\varepsilon}, \bar{x}^\varepsilon(t), \bar{u}(t)\right), \qquad \bar{x}^\varepsilon(0) = x_0.$$

Démonstration. On a

$$\frac{d}{ds}(\bar{x}^\varepsilon(s) - \bar{x}(s)) = g\left(\frac{s}{\varepsilon}, \bar{x}^\varepsilon(s), \bar{u}(s)\right) - g\left(\frac{s}{\varepsilon}, \bar{x}(s), \bar{u}(s)\right) + G^\varepsilon(s)$$

où

$$|G^\varepsilon(s)|_{L^2(0,T)} \leqq C\varepsilon^{K+1}.$$

Avec la condition initiale

$$\bar{x}^\varepsilon(0) - \bar{x}(0) = 0$$

on peut déduire aisément (36). □

Maintenant on peut énoncer le théorème 11.

THEOREME 11. *On a*

$$\left|\inf_{v(\cdot)} J^\varepsilon(v(\cdot)) - J^\varepsilon(\bar{u}(\cdot))\right| \leqq C\varepsilon^{2K+2}. \tag{87}$$

De plus si $u^\varepsilon(\cdot)$ est un contrôle meilleur que $\bar{u}$, i.e., si

$$J^\varepsilon(u^\varepsilon(\cdot)) \leqq J^\varepsilon(\bar{u}(\cdot)). \tag{88}$$

Alors on a

$$|u^\varepsilon - \bar{u}|_{L^2(0,T)}, |x^\varepsilon - \bar{x}|_{C(0,T)} \leqq C\varepsilon^{K+1}. \tag{89}$$

Démonstration. On peut appliquer (78) avec $v(\cdot) = \bar{u}(\cdot)$. D'après (86), on a

$$\left|J^\varepsilon(\bar{u}(\cdot)) - \int_0^T \left(H\left(\frac{t}{\varepsilon}, \bar{\sigma}(t)\right) - \bar{p}(t)\bar{x}(t)\right) dt - h(\bar{x}(T))\right| \leqq C\varepsilon^{2K+1}. \tag{90}$$

Soit maintenant v un contrôle meilleur que $\bar{u}$, en tenant compte (78), (90) on a

$$\begin{aligned} C\varepsilon^{2k+2} \geqq & \int_0^T \int_0^1 \int_0^1 \lambda H_{\beta\beta}\left(\frac{s}{\varepsilon}\bar{\beta}(t) + \lambda\mu\hat{\beta}, \bar{p}(t)\right)\hat{\beta}^2(t)\, d\lambda\, d\mu\, dt \\ & - C\varepsilon^{K+1}|\hat{x}(t)| - C\varepsilon^{K+1}\int_0^T |\hat{\beta}(t)|\, dt. \end{aligned} \tag{91}$$

Mais d'après (3.17), (3.18) on a

$$H_{\beta\beta}\left(\frac{t}{\varepsilon}, \bar{\beta}(t) + \lambda\mu\hat{\beta}, \bar{p}(t)\right)\hat{\beta}^2(t) \geqq \beta|Z(\lambda, \mu, t)|^2 \tag{92}$$

où

$$Z(\lambda, \mu, t) = \hat{u}(t) + H_{vv}^{-1}H_{vx}\left(\frac{t}{\varepsilon}\bar{\beta} + \lambda\mu\hat{\beta}, \bar{p}\right)\hat{x}(t). \tag{93}$$

Donc

$$C\varepsilon^{2k+2} \geqq \beta\int_0^T \int_0^1 \int_0^1 \lambda|Z(\lambda, \mu, t)|^2\, d\lambda\, d\mu\, dt - C\varepsilon^{K+1}|\hat{x}(t)| - C\varepsilon^{K+1}\int_0^T |\hat{\beta}(t)|\, dt.$$

D'autre part, on peut vérifier que

$$\max_{0 \leqq t \leqq T} |\hat{x}(t)| \leqq C\varepsilon^{k+1} + C\int_0^T \int_0^1 \int_0^1 |Z(\lambda, \mu, t)|\, d\lambda\, d\mu\, ds. \tag{94}$$

En effet on a

$$\frac{d\hat{x}}{dt} = g\left(\frac{t}{\varepsilon}, x^{\varepsilon}(t), v(t)\right) - g\left(\frac{t}{\varepsilon}, \bar{x}(t), \bar{u}(t)\right) + G(t), \qquad \hat{x}(0) = 0$$

où d'après (86)

$$\int_0^T |G(t)|\, dt \leqq c\varepsilon^{K+1}; \tag{95}$$

donc

$$\begin{aligned}\frac{d\hat{x}}{dt} &= A\hat{x}(t) + B\hat{u}(s) + G(t)\\ &= \bar{A}\hat{x}(t) + BZ(\lambda, \mu, t) + G(t)\end{aligned}$$

où

$$\bar{A} = A - BH_{vv}^{-1}H_{vx}\left(\frac{t}{\varepsilon}, \bar{\beta}(t) + \lambda\mu\hat{\beta}\right) \text{ est borné,}$$

ce qui avec (95) implique (94). D'après la définition (93) on a

$$\int_0^T |\hat{u}(t)|\, dt \leqq C \int_0^T |Z(\lambda, \mu, t)|\, dt + C\varepsilon^{K+1} \tag{96}$$

ce qui avec (91), (92) implique

$$\int_0^T \int_0^1 \int_0^1 \lambda |Z(\lambda, \mu, t)|^2\, d\lambda\, d\mu\, dt \leqq C\varepsilon^{2k+2}. \tag{97}$$

Encore de (93), (95) on a

$$|\hat{x}(t)|_{C(0,T)} \leqq C\varepsilon^{K+1}, \tag{98}$$

$$\int_0^T |\hat{u}(t)|\, dt \leqq C\varepsilon^{K+1}. \tag{99}$$

On a donc (89), ce qui avec (78) implique (87). □

5. Résolution asymptotique de l'équation de Bellman. On va étudier dans ce paragraphe la méthode des perturbations pour l'équation de Bellman

$$\frac{\partial \Phi^{\varepsilon}}{dt} + \inf_{v}\left[D\Phi^{\varepsilon} g\left(\frac{T-t}{\varepsilon}, x, v\right) + l\left(\frac{T-t}{\varepsilon}, x, v\right)\right] = 0, \qquad \Phi^{\varepsilon}(x, T) = h(x). \tag{100}$$

De cette équation on peut déduire le feedback optimal $u^{\varepsilon}(x, t)$, solution de

$$\begin{aligned}&\frac{\partial \Phi^{\varepsilon}}{dt} + \left(D\Phi^{\varepsilon} g\left(\frac{T-t}{\varepsilon}, x, u^{\varepsilon}\right)\right) + l\left(\frac{T-t}{\varepsilon}, x, u^{\varepsilon}\right) = 0,\\ &D\Phi^{\varepsilon} \cdot g_v\left(\frac{T-t}{\varepsilon}, x, u^{\varepsilon}\right) + l_v\left(\frac{T-t}{\varepsilon}, x, u^{\varepsilon}\right) = 0, \qquad \Phi^{\varepsilon}(T, x) = h(x).\end{aligned} \tag{101}$$

On va montrer que ce contrôle feedback peut être approché par le contrôle feedback du problème homogénéisé. On s'intéresse aussi à l'approximation de contrôle feedback d'ordre plus élevé.

5.1. Développement asymptotique de l'équation de Bellman. On va essayer de trouver un développement de Φ^ε, u^ε de la forme

$$\begin{aligned}\Phi^\varepsilon(x,t)&=\Phi_0(x,t)+\varepsilon\Phi_1\left(x,t,\frac{T-t}{\varepsilon}\right)+\cdots,\\ u^\varepsilon(x,t)&=u_0\left(x,t,\frac{T-t}{\varepsilon}\right)+\varepsilon u_1\left(x,t,\frac{T-t}{\varepsilon}\right)+\cdots,\end{aligned}\tag{102}$$

où $\Phi_i(x,t,\tau)$, $u_i(x,t,\tau)$ sont 1-périodiques par rapport à τ. D'après (100), (101) on a

$$-\frac{\partial\Phi_1}{\partial\tau}+\frac{\partial\Phi_0}{\partial t}+D\Phi_0 g(\tau,x,u_0(x,t,\tau))+l(\tau,x,u_0(x,t,\tau))=0,\qquad \Phi_0(x,T)=h(x),\tag{103}$$

$$D\Phi_0 g_v(\tau,x,u_0(x,t,\tau))+l_v(\cdot)=0,\tag{104}$$

$$-\frac{\partial\Phi_2}{\partial\tau}+\frac{\partial\Phi_1}{\partial t}+D\Phi_1 g(\tau,x,u_0(x,t,\tau))=0,\qquad \Phi_1(x,T,0)=0.\tag{105}$$

Généralement si on note

$$\begin{aligned}H_0&=D\Phi_0 g(\tau,x,v)+l(\tau,x,v),\\ H_i&=D\Phi_i g(\tau,x,v),\qquad i=1,2,\cdots,\end{aligned}$$

de façon similaire à ce qui a été fait dans § 4, on peut développer (100) comme

$$\begin{aligned}&\frac{\partial\Phi_{2l}}{\partial t}-\frac{\partial\Phi_{2l+1}}{\partial\tau}+\sum_{\langle l,\alpha\rangle}^{2l}\frac{1}{\alpha_1!\alpha_2!\cdots\alpha_l!}D_v^{(\alpha_1+\cdots+\alpha_l)}H_\alpha(\cdot)u_1^{\alpha_1}\cdots u_l^{\alpha_l}=0,\\ &\frac{\partial\Phi_{2l+1}}{\partial t}-\frac{\partial\Phi_{2l+2}}{\partial\tau}+\sum_{\langle l,\alpha\rangle}^{2l+1}\frac{1}{\alpha_1!\cdots\alpha_l!}D_v^{(\alpha_1+\cdots+\alpha_l)}H_{\alpha_0}(\cdot)u_1^{\alpha_1}\cdots u_l^{\alpha_l}=0,\end{aligned}\tag{106}$$

$$\Phi_l(x,T,0)=0,\qquad l=1,2,\cdots,K,\tag{107}$$

$$\sum_{\langle l,\alpha\rangle}^{l}\frac{1}{\alpha_1!\cdots\alpha_l!}D_v^{(1+\alpha_1+\cdots+\alpha_l)}H_{\alpha 0}(\cdot)u_1^{\alpha_1}\cdots u_l^{\alpha_l}=0\tag{108}$$

où $\sum_{\langle l,\alpha\rangle}^{j}$ désigne

$$\sum_{\substack{\alpha_0+\alpha_1+\alpha_2+\cdots+l\alpha_l=j,\\ \alpha_1,\cdots,\alpha_n\geqq 0}}.\tag{109}$$

5.2. Résolution et estimation de Φ_i et u_i. On va montrer comment résoudre Φ_i, u_i, $i=0,1,\cdots$, de façon récursive. On va résoudre et estimer spécifiquement Φ_0, (x,t), $u_0(x,t,x)$, $\Phi_1(x,t,\tau)$, les autres peuvent se traiter de façon analogue.

On considère tout d'abord (103). D'après la remarque 6 pour que $\Phi_1(x,t,\tau)$ soit périodique par rapport à τ, on a nécessairement

$$\begin{aligned}&\frac{\partial\Phi_0}{\partial t}+D\Phi_0\int_0^1 g(\tau,x,u_0(t,\tau,x))\,d\tau+\int_0^1 l(\tau,x,u_0(t,\tau,x))\,d\tau=0,\\ &\Phi_0(T,x)=h(x);\end{aligned}\tag{110}$$

on peut relier (103) et (104) à un problème d'optimalité

$$\begin{aligned}&\frac{dx}{ds}=\int_0^1 g(\tau,x(s),u(s,\tau))\,d\tau,\quad v(s,\tau)\in L^2([t,\tau]\times[0,1];\mathbb{R}^K),\quad x(t)=x,\\ &\Phi_0(x,t)=\inf_{v(\cdot)}\left[\int_t^T\int_0^1 l(\tau,x(s),v(s,\tau))\,ds\,d\tau+h(x(T))\right].\end{aligned}\tag{111}$$

Mais on sait que sous les hypothèses du théorème 5, avec $p=2K$, on a

$$|D_x^{(j)}\Phi_0(x,t)| \leqq C(1+|x|)^{2-j}, \qquad 0 \leqq j \leqq 4K+1, \tag{112}$$

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^i}\Phi_0(x,t)\right| \leqq C(1+|x|)^{2-j}, \qquad 0 \leqq i+j \leqq 4K+1, \tag{113}$$

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^i}u_0(t,\tau,x)\right| \leqq C(1+|x|)^{1-j}, \qquad 0 \leqq i+j \leqq 4K. \tag{114}$$

On va montrer comment on peut trouver $\Phi_1(x,t,\tau)$, $\Phi_2(x,t,\tau)$, et $u_1(x,t,\tau)$, les autres peuvent se déduire de la même manière, de façon récursive.

Ayant résolu Φ_0, u_0, on peut résoudre Φ_1 de la façon suivante. On définit

$$\Phi_1(\bar{x},t,\tau) = \int_0^\tau \left[\frac{\partial\Phi_0}{\partial t} + D\Phi_0 g(s,x,u_0(x,t,s)) + l(s,x,u_0)\right] ds. \tag{115}$$

D'après la remarque 6, $\bar{\Phi}_1$ est 1-périodique, c'est également une solution de (103). Mais toujours, d'après la remarque 6, Φ_1 et $\bar{\Phi}_1$ peuvent différer d'une constante.

On peut donc écrire

$$\Phi_1(x,t,\tau) = \bar{\Phi}_1(x,t,\tau) + \psi_1(x,t) \tag{116}$$

où $\psi_1(x,t)$ peut se résoudre par (105). En effet, on a

$$\frac{\partial\psi_1}{\partial t} + D_x\psi_1 \int_0^1 g(\tau,x,u_0(x,t,\tau))\,d\tau + \int_0^1 \left(D_x\bar{\Phi}_1 g + \frac{\partial\bar{\Phi}_1}{\partial t}\right) d\tau = 0, \qquad \psi_1(T,x) = 0. \tag{117}$$

C'est un type de l'équation de Hamilton-Jacobi (linéaire).

Grâce au (112)-(114) on sait que

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^i}\Phi_1(x,t,\tau)\right| \leqq C(1+|x|)^{2-j}, \qquad 0 \leqq i+j \leqq 4K-1. \tag{118}$$

Maintenant on peut trouver $u_1(x,t,\tau)$. Si on considère (108) en cas où $l=1$, on a

$$D\Phi_1 g_v(\tau,x,u_0(x,t,\tau)) + (D_x\Phi_0 g_{vv} + l_{vv})u_1 = 0. \tag{119}$$

On a donc

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^i}u_1\right| \leqq C(1+|x|)^{1-j}, \qquad 0 \leqq i+j \leqq 4K-2. \tag{120}$$

De la même façon on peut noter

$$\Phi_2(x,t,\tau) = \bar{\Phi}_2(x,t,\tau) + \psi_2(x,t)$$

où $\bar{\Phi}_2(x,t,\tau)$ peut se résoudre par (105):

$$\bar{\Phi}_2(x,t,\tau) = \int_0^\tau \left(\frac{\partial\Phi_1}{\partial t} + D\Phi_1 g(\tau,x,u_0(t,\tau,x))\right) d\tau$$

et $\psi_2(x,t)$ peut se résoudre par (106) en cas où $l=1$.

Généralement on peut résoudre Φ_l, u_l par (106)-(108), de façon analogue.

De plus on a

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^1}\Phi_l(t,\tau,x)\right| \leqq C(1+|x|)^{2-j}, \qquad 0 \leqq i+j+2l \leqq 4K+1,$$

$$\left|D_x^{(j)}\frac{\partial^{(i)}}{\partial t^1}u_l(t,\tau,x)\right| \leqq C(1+|x|)^{1-j}, \qquad 0 \leqq i+j+2l \leqq 4K, \quad l \leqq K. \tag{121}$$

5.3. Résultat de convergence. On note que

$$\bar{\Phi}^\varepsilon(x, s) = \Phi_0(x, s) + \varepsilon\Phi_1\left(x, s, \frac{T-s}{\varepsilon}\right) + \cdots + \varepsilon^{2K+1}\Phi_{2K+1}\left(x, s, \frac{T-s}{\varepsilon}\right),$$

$$\bar{u}^\varepsilon(x, s) = u_0\left(x, s, \frac{T-s}{\varepsilon}\right) + \cdots + \varepsilon^K u_K\left(x, s, \frac{T-s}{\varepsilon}\right).$$

D'après (106)-(108) et les estimations (121), on a

$$\left[\frac{\partial\bar{\Phi}^\varepsilon}{\partial s} + D\bar{\Phi}^\varepsilon g\left(\frac{T-s}{\varepsilon}, x, \bar{u}^\varepsilon\right) + l\left(\frac{T-s}{\varepsilon}, x, \bar{u}^\varepsilon\right)\right]_{\langle l\rangle} = 0, \qquad l = 0, 1, \cdots, 2K+1.$$

Donc

$$\begin{aligned}
&\left[\frac{\partial\bar{\Phi}^\varepsilon}{\partial s} + D_x\bar{\Phi}^\varepsilon(x, s)\cdot g\left(\frac{T-s}{\varepsilon}, x, \bar{u}^\varepsilon\right) + l\left(\frac{T-s}{\varepsilon}, x, \bar{u}^\varepsilon\right)\right] \\
&\qquad \leqq C\varepsilon^{2K+2}\left(1 + \sum_{i=1}^{k}|u_i|^2 + \sum_{i=1}^{2K+1}(|D_x\Phi_i|^2 + |x|^2)\right) \qquad (122) \\
&\qquad \leqq c\varepsilon^{2K+2}(1+|x|^2).
\end{aligned}$$

On a le théorème 12.

THEOREME 12. *On fait les hypothèses du théorème 5 avec $p = 2K$. Alors on a*

$$|\Phi^\varepsilon(x, t) - \bar{\Phi}^\varepsilon(x, t)| \leqq c\varepsilon^{2K+2}(1+|x|^2). \qquad (123)$$

Démonstration. On note $x^\varepsilon(t)$, $v(t)$, solution de

$$\frac{dx^\varepsilon}{ds} = g\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right), \qquad x^\varepsilon(t) = x. \qquad (124)$$

On peut limiter les contrôles admissibles par

$$\int_t^T |U(s)|^2\, ds \leqq C(1+|x|^2). \qquad (125)$$

Par conséquent

$$|x^\varepsilon(s)|_{C(t,T)} \leqq C(1+|x|). \qquad (126)$$

On considère

$$\begin{aligned}
&\frac{\partial\bar{\Phi}^\varepsilon}{\partial s} + D_x\bar{\Phi}^\varepsilon g\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right) + l\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right) \\
&\quad = \frac{\partial\bar{\Phi}^\varepsilon}{\partial s} + D_x\bar{\Phi}^\varepsilon g\left(\frac{s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon(x^\varepsilon(s), s)\right) + l\left(\frac{s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon(\cdot)\right) \\
&\qquad + \left(D_x\bar{\Phi}^\varepsilon g_v\left(\frac{s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon\right) + l_v\left(\frac{s}{\varepsilon}, x^\varepsilon, \bar{u}^\varepsilon\right)\right)v(s) \\
&\qquad + \int_0^1\int_0^1 \lambda\left(D_x\bar{\Phi}^\varepsilon g_{vv}\left(\frac{s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon + \lambda\mu\hat{u}\right)\right. \\
&\qquad \left. + l_{vv}\left(\frac{s}{\varepsilon}, x^\varepsilon(s)^*, \bar{u}^\varepsilon + \lambda\mu\hat{u}\right)\right)(v(s))^2\, d\lambda\, d\mu.
\end{aligned}$$

D'après (126) et (121) on a

$$\frac{\partial \bar{\Phi}^\varepsilon}{\partial s} + D_x \bar{\Phi}^\varepsilon \cdot g\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right) \geqq -C\varepsilon^{2K+2}(1+|x^\varepsilon(s)|^2+|v(s)|^2)$$

(127)
$$- l\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right),$$

$$\bar{\Phi}^\varepsilon(x^\varepsilon(T), T) = h(x^\varepsilon(T)).$$

En intégrant des deux côtés de (127), on a

$$\bar{\Phi}^\varepsilon(x, t) \leqq \int_t^T l\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right) ds + h(x^\varepsilon(T)) - C\varepsilon^{2K+2} \int_t^T (1+|x^\varepsilon(s)|^2+|v(s)|^2)\, ds$$

d'après (125), (126)

$$\leqq \int_t^T l\left(\frac{s}{\varepsilon}, x^\varepsilon(s), v(s)\right) ds + h(x^\varepsilon(T)) - C\varepsilon^{2K+2}(1+|x|^2).$$

En particulier, on a

(128)
$$\Phi^\varepsilon(x, t) - \bar{\Phi}^\varepsilon(x, t) \geqq C\varepsilon^{2K+2}(1+|x|^2).$$

L'inégalité contraire peut être démontrée de la façon suivante. On introduit un système de contrôle feedback

(129)
$$\frac{dx^\varepsilon}{ds} = g\left(\frac{T-s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon\left(x^\varepsilon(s), s, \frac{T-s}{\varepsilon}\right)\right), \qquad x^\varepsilon(t) = x.$$

D'après (121), on a

(130)
$$|x^\varepsilon(s)|_{c(t,T)}, \left|\bar{u}^\varepsilon\left(x^\varepsilon(s), s, \frac{T-s}{\varepsilon}\right)\right|_{L^2(t,T)} \leqq C(1+|x|^2).$$

D'autre part, de (100) on a

$$\frac{\partial \Phi^\varepsilon}{\partial t} + D\Phi^\varepsilon \cdot g\left(\frac{T-s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon\right) + l\left(\frac{T-s}{\varepsilon}, x^\varepsilon(s), \bar{u}^\varepsilon\right) \geqq 0.$$

Ce qui avec (122) implique

(131)
$$\frac{\partial(\Phi^\varepsilon - \bar{\Phi}^\varepsilon)}{\partial s} + D(\Phi^\varepsilon - \Phi^{\bar{\varepsilon}}) \cdot g\left(\frac{s}{\varepsilon}, x^\varepsilon, \bar{u}^\varepsilon\right) \geqq -C\varepsilon^{2K+2}(1+|x|^2+|\bar{u}^\varepsilon|^2).$$

De (107)

(132)
$$\Phi^\varepsilon(x, T) - \bar{\Phi}^\varepsilon(x, T) = 0.$$

On intègre des deux côtés de (131), avec (130), (132).

On a finalement

(133)
$$\Phi^\varepsilon(t, x) - \bar{\Phi}^\varepsilon(t, x) \leqq C\varepsilon^{2K+2}(1+|x|^2).$$

Ce qui avec (128) complète la démonstration. □

6. Exemple. On suppose $0 < \theta < 1$, et

$$g(\tau, x, v) = \begin{cases} g_1(x, v), & 0 \leqq \tau < \theta, \\ g_2(x, v), & \theta \leqq \tau < 1, \end{cases}$$

$$l(\tau, x, v) = \begin{cases} l_1(x, v), & 0 \leqq \tau < \theta, \\ l_2(x, v), & \theta \leqq \tau < 1. \end{cases}$$

On cherche le problème limité du problème

$$\frac{\partial \Phi^\varepsilon}{\partial t}+\inf_v\left[D\Phi^\varepsilon g\left(\frac{t}{\varepsilon}, x, v\right)+l\left(\frac{t}{\varepsilon}, x, v\right)\right]=0, \qquad \Phi^\varepsilon(x, T)=h(x). \tag{134}$$

C'est donc l'équation de Bellman homogénéisée

$$\frac{\partial \Phi}{\partial t}+\inf_{v(\tau)}\left[\int_0^1 (D\Phi \cdot g(\tau, x, v(\tau))+l(\tau, x, v(\tau)))\, d\tau\right]=0, \qquad \Phi(x, T)=h(x). \tag{135}$$

Donc

$$\frac{\partial \Phi}{\partial t}+\inf_{v(\tau)}\left[\int_0^\theta (D\Phi \cdot g_1(x, v(\tau))+l_1(x, v(\tau)))\, d\tau + \int_\theta^1 (D\Phi \cdot g_2(x, v(\tau))+l_2(x, v(\tau)))\, d\tau\right]=0.$$

Cela est équivalent à

$$\frac{\partial \Phi}{\partial t}+\inf_{v_1, v_2}\left[D\Phi \cdot (\theta \cdot g_1(x, v_1)+(1-\theta)g_2(x, v_2))+\theta l_1(x, v_1)+(1-\theta)l_2(x, v_2)\right]=0.$$

Mais on sait que c'est l'équation de Bellman du problème d'optimisation

$$\begin{aligned}\frac{dx}{dt}&=(\theta g_1(x, t), v_1(t))+(1-\theta)g_2(x(t), v_2(t))\\ &\cdot \min_{v_1(t), v_2(t)}\left[\int_0^T (\theta l_1(x(t), v_1(t))+(1-\theta)l_2(x(t), v_2(t))\, dt+h(x(t))\right].\end{aligned} \tag{136}$$

REFERENCES

[1] A. BENSOUSSAN, *Perturbation Methods in Optimal Control*, Dunod, Gauthier-Villars, Paris, 1988.

[2] ———, *Singular perturbations for deterministic control problems*, in Singular Perturbations and Asymptotic Analysis in Control Systems, P. Kokotovic, A. Bensoussan, and G. Blankenship, eds., Lecture Notes in Control and Inform. Sci. 90, Springer-Verlag, Berlin, New York, 1986, pp. 9-171.

[3] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Asymptotic Methods for Periodic Structures*, North-Holland, Amsterdam, New York, 1978.

[4] W. FLEMING AND R. RISHEL, *Optimal Deterministic and Stochastic Control*, Springer-Verlag, Berlin, New York, 1978.

[5] P. V. KOKOTOVIC, *Applications of singular perturbation techniques to control problems*, SIAM Rev., 26 (1984), pp. 501-550.

[6] J. L. LIONS, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Math. 323, Springer-Verlag, Berlin, New York, 1973.

[7] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[8] S.-G. PENG, *Etude de perturbations singulières en contrôle optimal deterministe*, Thèse de troisième cycle, Université de Paris-IX, 1985.

[9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, Wiley-Interscience, New York, 1962.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 697-717, July 1989

002

OPTIMAL DESIGN FOR HEAT CONDUCTION PROBLEMS WITH HYSTERESIS*

MARTIN BROKATE† AND AVNER FRIEDMAN‡

Abstract. A model of controlling the temperature in heat conduction is considered; the control mechanism is accomplished by means of a hysteresis operator that depends on some parameters. A cost functional is introduced that depends on these parameters and then properties for the optimal parameters are derived.

Key words. heat equation, hysteresis, thermostat control, optimal design, Preisach operator

AMS(MOS) subject classifications. 93C20, 49A36, 49B22

0. Introduction. In this paper we consider optimal design of a thermostat-like feedback mechanism in order to regulate heat flux. Assume that the temperature $u = u(x, t)$ satisfies

$$
\begin{aligned}
&u_t(x, t) = u_{xx}(x, t) && \text{in } (0, a)\times(0, T),\\
&u(x, 0) = u_0(x) && \text{in } (0, a),\\
&-u_x(0, t) = k(t) && \text{in } (0, T),\\
&u_x(a, t) + u(a, t) = 1 - f(t) && \text{in } (0, T).
\end{aligned}
\tag{0.1}
$$

Here u_0 and k are known functions with values in $[0, 1]$, and f is determined by a feedback from the temperature $u(0, \cdot)$ at the left end. For this feedback, let us momentarily consider the standard thermostat described graphically in Fig. 1.

Figure 1 means that the "output" f switches from 0 to 1 if the "input" $\tilde{u}(t) = u(0, t)$ takes on the threshold value ρ_2, and from 1 to 0 at the threshold value ρ_1. Assume

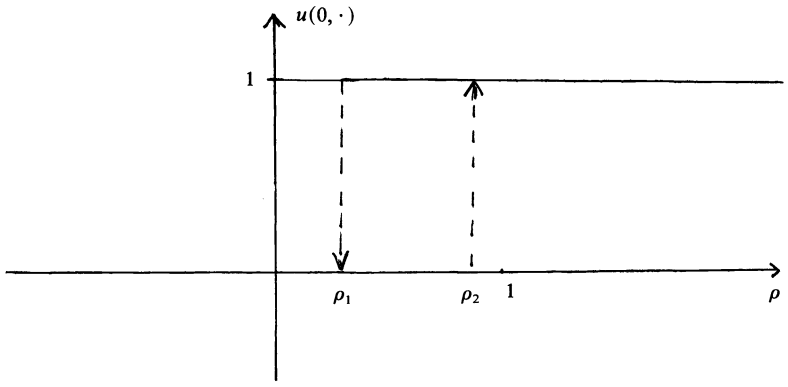

FIG. 1

* Received by the editors February 9, 1988; accepted for publication (in revised form) December 5, 1988. This work was partially supported by National Science Foundation grant DMS-86-12880 and Deutsche Forschungsgemeinschaft.

† Universität Kaiserslautern, Kaiserslautern, Federal Republic of Germany.

‡ Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

for the moment that any choice $\rho = (\rho_1, \rho_2)$ of threshold values $\rho_1 < \rho_2$ defines an operator W_ρ such that the feedback

$$f(t) = [W_\rho(u(0, \cdot))](t) \tag{0.2}$$

together with (0.1) uniquely determine the temperature evolution $u = u(x, t)$. Feedback (0.2) means that there is a heat inflow at $x = a$ if temperature at $x = 0$ is low and a heat outflow at $x = a$ if temperature at $x = 0$ is high. We now can pose the question: How should we choose the threshold $\rho = (\rho_1, \rho_2)$ to achieve some prescribed goal? For example, we might want to minimize the expression

$$\int_0^T (u(0, t) - g(t))^2 \, dt, \tag{0.3}$$

given the function g. This is an optimal design problem. (Note that the choice of the initial condition u_0 should not matter very much if T is large.) As a second question we may ask whether thermostat control is optimal within a larger class of feedback mechanisms

$$f(t) = W(u(0, \cdot)), e, \tag{0.4}$$

where W now stands for a suitable family of input-output-operators parametrized by some design parameter e.

The present paper contributes to these questions from the viewpoint of first-order necessary optimality conditions. Its main line of reasoning is the following: For the thermostat feedback (0.2), the map $(\rho_1, \rho_2) \mapsto u = u(x, t)$ is discontinuous. We therefore replace the single thermostat by a family of thermostats, i.e., we use (0.4) with W being the Preisach operator, defined by

$$W(\tilde{u}, \mu)(t) = \int_\Delta (W_\rho \tilde{u})(t) \, d\mu(\rho) \tag{0.5}$$

where $\Delta = \{(\rho_1, \rho_2): 0 \leqq \rho_1 < \rho_2 \leqq 1\}$ and μ is a finite Borel measure on Δ. If μ is the Dirac measure concentrated in $\rho \in \Delta$, we get back to the thermostat, so looking for an optimal design μ can be viewed as a relaxation of the thermostat design problem. On the other hand, if the measure μ has a bounded density e, the map $\tilde{u} \mapsto W(\tilde{u}, \mu)$ is Lipschitz continuous on $C[0, T]$, but not differentiable. Such measures thus regularize the problem while retaining the hysteresis, which is the main feature of the thermostat, but still we cannot apply first-order optimality conditions directly. Therefore, we approximate W by smooth operators W_ε so as to get a smooth approximating optimal design problem for which first-order optimality conditions can be readily developed. We next restrict the set of admissible measures μ to a set parametrized by $v = (\eta, \rho_1, \rho_2) \in \mathbb{R}^3$ such that μ_v is a smooth approximation of the Dirac measure at (ρ_1, ρ_2) with support in the η-ball around (ρ_1, ρ_2). For this restricted optimal design problem, we are finally able to conclude some properties of the optimal parameters $(\eta^*, \rho_1^*, \rho_2^*)$ in the limit $\varepsilon \to 0$, i.e., with the feedback (0.4) and (0.5); unfortunately, we can only handle the case where the boundary condition in (0.1) at $x = a$ is replaced by

$$u_x(a, t) + u(a, t) = f(t) \quad \text{in } (0, T). \tag{0.6}$$

Having presented the main steps of this paper, some historical comments are in order. The Preisach operator W was introduced in Preisach [9] to model hysteresis phenomena in ferromagnetism. Mathematical treatment was given in Krasnosel'skii and Pokrovskii [8] (see also [7]) and by Visintin [10], and was recently extended in Brokate and Visintin [3]. So far, first-order conditions in optimization problems

involving the Preisach operator have not been discussed in the literature. For the standard optimal control problem for an ODE system, coupled with a hysteresis operator W being less general than the Preisach operator, a smooth approximation W_ε has been given in Brokate [1] (see also [2]), and the limit procedure $\varepsilon \to 0$ has been carried out successfully in the first-order necessary conditions. The approximation W_ε adopted in the present paper is, in fact, a generalization of the one in [1], based on the description of the Preisach operator in [3].

Since the optimal design problem treated here is smooth except for the Lipschitz continuous operator W, we are tempted to apply nonsmooth optimization theory directly, thereby avoiding the ε-problem and the limit procedure. However, the feedback equation (0.4) is a major obstacle for this approach, as it represents a nonsmooth equality constraint with infinite dimensional range.

The structure of the paper is as follows. In § 1 we introduce the control problem for heat conduction. The cost functional represents the objective of achieving "comfortable temperatures" at "low cost." We shall also derive necessary conditions for the optimal control v_* which, however, are merely formal (since W is not differentiable). In § 2 we develop some aspects of the theory of Preisach operators that are needed in the sequel; this is based on ideas and facts from Brokate and Visintin [3]. In § 3 we introduce the smooth approximations W_ε to the Preisach operator W. In § 4 the corresponding optimal control problem (P_ε) is introduced, and its solutions v_ε are shown to converge to v_* as $\varepsilon \to 0$. In § 5 we derive necessary optimality conditions for v_ε and in § 6 we establish some estimates (uniformly in ε); these are translated, in § 7, into specific properties of v_ε and, subsequently, also of v_*. Some generalizations are mentioned in § 8.

1. The optimal design problem. Consider the heat equation

$$u_t = u_{xx} \quad \text{in } Q_T = \{(x, t);\ 0 < x < a,\ 0 < t < T\} \tag{1.1}$$

with initial condition

$$u(x, 0) = u_0(x), \qquad 0 < x < a \tag{1.2}$$

and boundary conditions

$$-u_x(0, t) = k(t), \qquad 0 < t < T, \tag{1.3}$$

$$u_x(a, t) + u(a, t) = W(u(0, \cdot), Bv)(t), \qquad 0 < t < T. \tag{1.4}$$

Here $u_0 \in C[0, a]$ and $k \in C[0, T]$ are given functions, $W = W(u, e)$ is the Preisach hysteresis operator to be defined in detail in § 2, and e represents the density of the measure μ in (0.5). The variable v denotes the design parameter to be chosen,

$$v \in K \subset V$$

represents its admissible range, V being a Banach space (later on to be taken as $\mathbb{R}^3$);

$$B: V \to L^\infty(\Delta)$$

is a continuously (Fréchet-)differentiable map where

$$\Delta = \{(\rho_1, \rho_2):\ 0 \leqq \rho_1 < \rho_2 \leqq 1\}.$$

It is easy to prove that for any $v \in K$ there exists a unique solution $u = u(v)$ of (1.1)-(1.4) (see § 4). We introduce the cost functional

$$J(v) = \int_0^T \Phi(u(0, t), t)\, dt + h(v) \tag{1.5}$$

where $u = u(v)$; Φ and h are given continuously differentiable functions. For example, as in (0.3) we may set

$$\Phi(u, t) = (u - g(t))^2,$$

where g is a given continuous function.

Consider the problem:

Problem (P). Find $v_* \in K$ such that

$$J(v_*) = \min_{v \in K} J(v).$$

This problem represents a heat conduction model with a boundary feedback (1.4), chosen so as to be a relaxation of the standard thermostat feeedback for reasons explained in the Introduction. (If we set the right-hand side of (1.4) equal to $1 - W$ in order to conform to the thermostat (compare (0.1) and (0.3)), all the results of this paper except Lemma 6.2 and Theorem 7.1 remain true with obvious modifications.) The feedback can be adjusted with the control variable $v \in K$. The expense of heating (or cooling) at $x = a$ is measured by $h(v)$ and the goal of achieving desirable temperatures is measured by $\int_0^T \Phi(u(0, t), t)\, dt$. Various other functionals, expressing the goal of "desirable temperatures," can also be chosen; see § 8.

It is easy to prove (see § 4; cf. also [6]) that this problem has at least one solution v_*. Set $u_* = u(v_*)$ and introduce the sets

$$K^{\pm}(v) = \{w \in L^{\infty}(\Delta)\colon v \pm \lambda w \in K \text{ for all sufficiently small } \lambda,\ \lambda > 0\}.$$

Then

$$J(v_* + \lambda w) \geqq J(v_*) \quad \forall w \in K^{+}(v_*) \tag{1.6}$$

for any small λ, $\lambda > 0$. Assuming

$$u(v_* + \lambda w) = u_* + \lambda z + o(\lambda)$$

we get from (1.6), after letting $\lambda \to 0$,

$$\int_0^T \Phi_u(u_*(0, t), t) z(0, t)\, dt + Dh(v_*)w \geqq 0 \quad \forall w \in K^{+}(v_*). \tag{1.7}$$

Here and in the following, D denotes (Fréchet-) derivative and D_e, D_u, etc., denote partial derivatives $\Phi_u = D_u\Phi$. The function z in (1.7) satisfies (formally only, since W is not differentiable)

$$\begin{aligned}
&z_t = z_{xx} \quad \text{in } Q_T,\\
&z_x(0, t) = 0, \qquad 0 < t < T,\\
&z_x(a, t) + z(a, t) = [D_u W(u_*(0, \cdot), Bv_*) z(0, \cdot)](t)\\
&\qquad\qquad + D_e W(u_*(0, \cdot), Bv_*) DB(v_*) \cdot w](t), \qquad 0 < t < T,\\
&z(x, 0) = 0, \qquad 0 < x < a.
\end{aligned} \tag{1.8}$$

We shall use the abbreviation

$$D_u W = D_u W(u_*(0, \cdot), Bv_*), \qquad D_e W = D_e W(u_*(0, \cdot), Bv_*) DBv_*$$

and introduce the adjoint operator $D_u W^*$ by

$$\int_0^T p(t)(D_u W z)(t)\, dt = \int_0^T z(t)(D_u W^* p)(t)\, dt \quad \forall p, z \in C[0, T]. \tag{1.9}$$

Let p be the solution of the "adjoint problem":

$$
\begin{aligned}
&p_t + p_{xx} = 0 \quad \text{in } Q_T,\\
&p(x, T) = 0, \qquad 0 < x < a,\\
&-p_x(0, t) = [D_u W^* p(a, \cdot)](t) + \Phi_u(u_*(0, t), t), \qquad 0 < t < T,\\
&p_x(a, t) + p(a, t) = 0, \qquad 0 < t < T.
\end{aligned} \tag{1.10}
$$

We easily compute, using (1.8),

$$
\begin{aligned}
0 &= \int\int_{Q_T} p(z_t - z_{xx}) + \int\int_{Q_T} z(p_t + p_{xx})\\
&= \int_0^T (-pz_x + zp_x)\, dt \Big|_{x=0}^{x=a}\\
&= -\int_0^T p(z_x + z)\, dt \Big|^{x=a} + \int_0^T pz\, dt \Big|^{x=a} + \int_0^T p_x z\, dt \Big|_{x=0}^{x=a}\\
&= -\int_0^T z(0, t)[(D_u W^* p(a, \cdot))(t) + p_x(0, t)]\, dt\\
&\qquad + \int_0^T z(a, t)[p(a, t) + p_x(a, t)]\, dt - \int_0^T p(a, t)(D_e W \cdot w)(t)\, dt
\end{aligned}
$$

and therefore, by the last two conditions in (1.10),

$$
\int_0^T z(0, t)\Phi_u(u_*(0, t), t)\, dt = \int_0^T p(a, t)(D_e W \cdot w)(t)\, dt. \tag{1.11}
$$

Substituting this into (1.7), we obtain the optimality conditions:

$$
\int_0^T p(a, t)(D_e W \cdot w)(t) + Dh(v_*)w \geqq 0 \quad \forall w \in K^+(v_*). \tag{1.12}
$$

Similarly we get the reverse inequality for all $w \in K^-(v_*)$.

Since W is not differentiable, these conditions do not actually make any sense; what we will do later is approximate W by smooth operators W_ε and problem (P) by problems (P_ε) corresponding to W_ε, and then derive, by the above method, rigorous optimality conditions.

2. The Preisach operator. The standard thermostat as depicted in Fig. 1 can be formalized as an operator $W_\rho : C[0, T] \to L^\infty(0, T)$ mapping any continuous function $u = u(t)$ to a function $y_\rho = y_\rho(t)$ that takes on the values 0 and 1 only, according to the switching rule indicated by the arrows in the figure. Since we will not use this approach, we refer to [10] and [3] for definition and basic properties of W_ρ. We remark, however, that W_ρ is discontinuous, no matter how the norms are chosen.

The Preisach operator W is usually defined as

$$
W(u, e)(t) = \int_\Delta (W_\rho u)(t) e(\rho)\, d\rho \tag{2.1}
$$

with $e \in L^\infty(\Delta)$, Δ being some subset of $\{\rho_1 < \rho_2\}$ that we specify here as

$$
\Delta = \{\rho = (\rho_1, \rho_2) : 0 \leqq \rho_1 < \rho_2 \leqq 1\},
$$

so that W represents an average over a continuous family of thermostats with respect to the measure $d\mu = e\, d\rho$. Let us fix $e \in L^\infty(\Delta)$. To study the time evolution of $W(u, e)$ for any $u \in C[0, T]$ means to study the time evolution of the sets (whose disjoint union is Δ)

$$A^+(t) = \{\rho: \rho \in \Delta, (W_\rho u)(t) = 0\}, \qquad A^-(t) = \{\rho: \rho \in \Delta, (W_\rho u)(t) = 1\} \tag{2.2}$$

for some initial configurations $A^+(0)$, $A^-(0)$ of the values of the thermostats that we assume to be given (later we will always set $A^-(0) = \varnothing$). It turns out that the boundary

$$B(t) = \partial A^+(t) \cap \partial A(t), \qquad t > 0, \tag{2.3}$$

is either empty (in case one of the sets is empty) or defines a line in the (ρ_1, ρ_2)-plane that separates Δ and consists of horizontal and vertical segments, the number of corner points being either finite or infinite with a limit point on the main diagonal [8], [10], if $B(0)$ already has that property; see Fig. 2. (The fastest way to believe this is to draw pictures for simple piecewise linear functions u.)

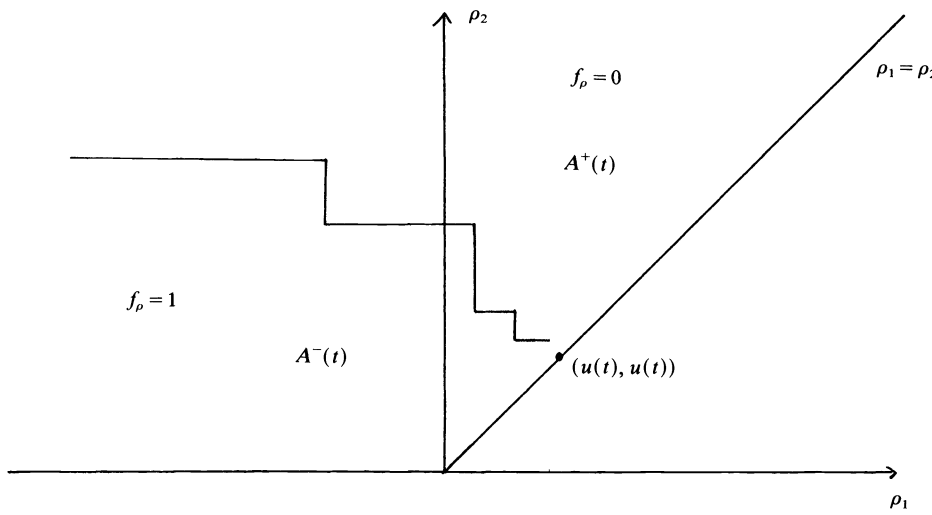

FIG. 2

Also, knowledge of $B(t)$ is all that is needed to determine $W(u, e)$ on $[t, T]$ from u on $[t, T]$, since it fixes the values of all thermostats except on the set $B(t)$ itself that has μ-measure zero. We may therefore interpret $B(t)$ as the internal state of a system whose input-output behavior is described by the Preisach operator (2.1) with fixed density e. Actually it is much more convenient for analysis to use such boundary curves as a basic description of the Preisach operator W, instead of the individual thermostats W_ρ. We only have to change coordinates to

$$r = \frac{\rho_2 + \rho_1}{2}, \qquad s = \frac{\rho_2 - \rho_1}{2}, \tag{2.4}$$

representing average and separation of the switching thresholds; then for any $t > 0$ the boundary curve $B(t)$ can be written as the graph of a Lipschitz function with slope ± 1, henceforth denoted by $\psi(t)$,

$$B(t) = \{(r, s) \in \tilde{\Delta}: s = \psi(t) r, r \geqq 0\},$$

where $\tilde{\Delta}$ is the image of Δ under the mapping (2.4), so that the internal state is now described by a function

$$\psi : [0, T] \to \mathrm{Lip}\,[0, \infty),$$

the values $\psi(t)r$ for $r \geqq 1$ being irrelevant for our specific choice of Δ. It is clear from (2.1)–(2.4) and Fig. 2 that, once the function ψ is known for a given $u \in C[0, T]$, we immediately obtain

$$W(u, e)(t) = \int_0^\infty \int_{-\infty}^{\psi(t)r} (\tilde{T}e)(r, s)\, ds\, dr,$$

where $\tilde{T}$ represents the transformation according to (2.4), extended by setting $\tilde{T}e(r, s) = 0$ for $(r, s) \notin \tilde{T}(\Delta)$.

This approach, investigated at length in [3], is of interest here mainly because we can readily obtain a smooth approximation W_ε of W from it. To make the present paper reasonably self-contained, we repeat some material from [3] as far as it is essential for the definition of W_ε.

From a rather informal discussion, we now switch to a formal one.

The spaces that will be used to define the internal states are the following:

$\tilde{\Psi}_0 = \{\phi : \phi \in C[0, \infty),\ \phi$ has compact support$\}$,

$\tilde{\Psi}_0$ is provided with sup-norm and its closure will be denoted by Ψ_0;

$\Psi_1 = \{\phi \in \Psi_0,\ \phi$ is Lipschitz continuous with Lipschitz constant $\leqq 1\}$.

We want to define a map

$$F : C[0, T] \times \Psi_1 \to C([0, T]; \Psi_1)$$

such that

$$\psi(t) = F(u, \psi_0)(t)$$

is the internal state at time t with respect to the (r, s)-coordinates, given an input u and an initial internal state ψ_0. The initial state ψ_0 fixes the values of the thermostats at time $t = 0$; here $\psi_0 \equiv 0$ corresponds to $A^-(0) = \varnothing$ so that all thermostats are off, whereas $\psi_0 \equiv 1$ corresponds to $A^-(0) = \Delta$.

LEMMA 2.1. *Define* $g : \mathbb{R}^3 \to \mathbb{R}$ *by*

$$g(x, p, r) = \min\{x + r, \max\{x - r, p\}\} \tag{2.5}$$

and $G : \mathbb{R} \times \Psi_0 \to \Psi_0$ *by*

$$G(x, \phi)(r) = g(x, \phi(r), r). \tag{2.6}$$

Then for all arguments we have

$$\|G(x_1, \phi_1) - G(x_2, \phi_2)\|_\infty \leqq \max\{|x_1 - x_2|, \|\phi_1 - \phi_2\|_\infty\}, \tag{2.7}$$

$$G : \mathbb{R} \times \Psi_1 \to \Psi_1, \tag{2.8}$$

$$G(x_1, \phi_1) \leqq G(x_2, \phi_2) \quad \text{if } x_1 \leqq x_2, \quad \phi_1 \leqq \phi_2. \tag{2.9}$$

Proof. Indeed (2.7) follows from

$$|g(x_1, p_1, r_1) - g(x_2, p_2, r_2)| \leqq \max\{|p_1 - p_2|, |x_1 - x_2| + |r_1 - r_2|\}, \tag{2.10}$$

(2.8) follows from (2.7), and (2.9) follows immediately from (2.6) and the appropriate monotonic behavior of g.

We may easily verify that, for monotone inputs $u \in C[0, T]$,

$$\psi(t) = G(u(t), \psi_0)$$

gives the correct boundary curve $B(t)$, if we choose ψ_0 corresponding to $B(0)$.

Denote by $C_{pm}[0, T]$ the set of functions $u \in C[0, T]$ that are piecewise monotone. We define

$$F: C_{pm}[0, T] \times \Psi_1 \to C([0, T]; \Psi_1)$$

by

$$\begin{aligned} &F(u, \psi_0)(0) = G(u(0), \psi_0), \\ &F(u, \psi_0)(t) = G(u(t), F(u, \psi_0)(t_i)) \quad \text{if } t \in [t_i, t_{i+1}], \end{aligned} \tag{2.11}$$

where $[t_j, t_{j+1}]$ are the intervals in which u is monotone.

LEMMA 2.2. *The function F satisfies*

$$\|F(u_1, \psi_{01}) - F(u_2, \psi_{02})\|_{C([0,T];\Psi_1)} \leqq \max \{\|u_1 - u_2\|_\infty, \|\psi_{01} - y_{02}\|_\infty\}. \tag{2.12}$$

Indeed this can be seen by induction on the monotonicity intervals of u_1, u_2.

From (2.12) with $\psi_{01} = \psi_{02}$ it follows that F can be extended into a Lipschitz continuous map with Lipschitz coefficient 1,

$$F: C[0, T] \times \Psi_1 \to C([0, T]; \Psi_1).$$

From (2.9), (2.11) we also deduce that

$$F(u_1, \psi_0) \leqq F(u_2, \psi_0) \quad \text{if } u_1 \leqq u_2. \tag{2.13}$$

Given any input $u \in C[0, T]$, we now define the internal state $\psi: [0, T] \to \Psi_1$ as

$$\psi(t) = F(u, \psi_0)(t). \tag{2.14}$$

For any $\phi \in \Psi_0$ we define the state-output map

$$E(\phi, e) = \int_0^\infty \int_{-\infty}^{\phi(r)} e(r, s)\, ds\, dr \tag{2.15}$$

where

$$e \in L^\infty(\mathbb{R}_+ \times \mathbb{R}), \qquad \operatorname{supp} e \subset \tilde{\Delta}$$

where $\tilde{\Delta}$ is the ρ-set Δ in the (r, s)-coordinates. Then

$$\begin{aligned} |E(\phi_1, e) - E(\phi_2, e)| &\leqq \int_0^1 \left| \int_{\phi_2(r)}^{\phi_1(r)} |e(r, s)|\, ds \right| dr \\ &\leqq \|\phi_1 - \phi_2\|_\infty \|e\|_\infty \end{aligned}$$

and

$$E(\phi, \cdot): L^\infty(\mathbb{R}_+ \times \mathbb{R}) \to \mathbb{R}$$

is linear continuous with

$$|E(\phi, e)| \leqq \|e\|_{L^1(\tilde{\Delta})} \leqq \|e\|_\infty \cdot \operatorname{meas}(\tilde{\Delta}).$$

We can verify that the Preisach operator as defined in (2.1) for the measure μ with $d\mu = e\, d\rho_1\, d\rho_2$ is precisely the operator

$$\begin{aligned} W(u, e)(t) &= E(\psi(t), \tilde{T}e) \\ &= E(F(u, \psi_0)(t), \tilde{T}e) \end{aligned} \tag{2.16}$$

where $\tilde{T}: L^\infty(\Delta) \to L^\infty(\tilde{\Delta})$ is defined by

$$(\tilde{T}e)(r, s) = e(s - r, s + r).$$

Combining properties of F and (2.16) we get Lemma 2.3.

LEMMA 2.3. *The operator W satisfies*

$$W: C[0, T]\times L^{\infty}(\Delta)\to C[0, T], \tag{2.17}$$

$$W(u, \cdot): L^{\infty}(\Delta)\to C[0, T] \text{ is a linear continuous operator,} \tag{2.18}$$

$$\|W(u, e)\|_{\infty}\leqq \text{meas}\,(\tilde{\Delta})\cdot\|e\|_{\infty}, \tag{2.19}$$

$$\|W(u_1, e)-W(u_2, e)\|\leqq\|u_1-u_2\|_{\infty}\|e\|_{\infty}. \tag{2.20}$$

If the input u is Lipschitz continuous then

$$\|F(u, \psi_0)(t)-F(u, \psi_0)(t')\|\leqq\|\dot{u}\|_{\infty}|t-t'| \quad \forall t, t'\in[0, T], \tag{2.21}$$

$\dot{u}$ denoting time derivative of u, as seen by considering first the case of piecewise monotone u's. Consequently,

$$|W(u, e)(t)-W(u, e)(t')|\leqq\|e\|_{\infty}\|\dot{u}\|_{\infty}|t-t'|. \tag{2.22}$$

Remark 2.1. In the sequel we shall continue to consider only the typical case of initial state $\psi_0\equiv 0$. However all the definitions and properties easily extend to the case of any initial state.

3. Regularization of the Preisach operator. In [1], a regularization of the so-called hysteresis operator of first kind has been developed. In this section, we extend this approach to the Preisach operator. Let $j\in C^{\infty}(\mathbb{R}^3)$, $j\geqq 0$ with support in the unit ball $\{|X|\leqq 1\}$, $\int j(X)\,dX=1$. For any small $\varepsilon>0$ set

$$j_{\varepsilon}(X)=\frac{1}{\varepsilon^3}j\left(\frac{X}{\varepsilon}\right)$$

and define the mollifier of the function g introduced in (2.5) by

$$g_{\varepsilon}(X)=(j_{\varepsilon}*g)(X)=\int_{\mathbb{R}^3}j_{\varepsilon}(X-Y)g(Y)\,dY, \qquad X\in\mathbb{R}^3.$$

Analogously to (2.6) we define $G_{\varepsilon}:\mathbb{R}\times\Psi_0\to\Psi_0$ by

$$G_{\varepsilon}(x, \phi)(r)=g_{\varepsilon}(x, \phi(r), r).$$

Then

$$g_{\varepsilon}\in C^{\infty}(\mathbb{R}^3; \mathbb{R}), \qquad G_{\varepsilon}\in C^{\infty}(\mathbb{R}\times\Psi_0; \Psi_0).$$

LEMMA 3.1. *The function G_{ε} satisfies, for every x_i, x in $\mathbb{R}$ and ϕ_i, ϕ in Ψ_0,*

$$\|G_{\varepsilon}(x_1, \phi_1)-G_{\varepsilon}(x_2, \phi_2)\|_{\infty}\leqq\max\{|x_1-x_2|, \|\phi_1-\phi_2\|_{\infty}\}, \tag{3.1}$$

$$G_{\varepsilon}(x_1, \phi_1)\leqq G_{\varepsilon}(x_2, \phi_2) \quad \text{if } x_1\leqq x_2,\ \phi_1\leqq\phi_2, \tag{3.2}$$

$$\|G_{\varepsilon}(x, \phi)-G(x, \phi)\|_{\infty}\leqq\varepsilon. \tag{3.3}$$

Proof. The inequalities (3.1), (3.2) follow easily from (2.7), (2.9), respectively. Next

$$\begin{aligned}
&|G_{\varepsilon}(x, \phi)(r)-G(x, \phi)(r)|\\
&\quad=\left|\int_{\mathbb{R}^3}j_{\varepsilon}(\xi, q, s)[g(x-\xi, \phi(r)-q, r-s)-g(x, \phi(r), s)]\,d\xi\,dq\,ds\right|\\
&\quad\leqq\left\{\int_{\mathbb{R}^3}j_{\varepsilon}(\xi, q, s)\,d\xi\,dq\,ds\right\}\varepsilon=\varepsilon \quad \text{by (2.10)}
\end{aligned}$$

and (3.3) follows.

To define F_ε we divide the interval $[0, T]$ by equidistant points $t_i = ih$ and then choose ε by $\sqrt{\varepsilon} = h$. We then define

$$F_\varepsilon : C[0, T] \times \Psi_0 \to \prod_i (C[t_i, t_{i+1}]: \Psi_0) \subseteq L^\infty([0, T]; \Psi_0)$$

by

$$\begin{aligned} &F_\varepsilon(u, \psi_0)(0) = G_\varepsilon(u(0), \psi_0), \qquad t \in [0, t_1], \\ &F_\varepsilon(u, \psi)(t) = G_\varepsilon(u(t), F_\varepsilon(u, \psi_0)(t_i - 0)), \qquad t \in [t_i, t_{i+1}]. \end{aligned} \tag{3.4}$$

LEMMA 3.2. *The function F_ε satisfies the following*:

$$\|F_\varepsilon(u_1, \psi_0) - F_\varepsilon(u_2, \psi_0)\|_{L^\infty(0,T;\Psi_0)} \leqq \|u_1 - u_2\|_\infty, \tag{3.5}$$

$$F_\varepsilon(u_1, \psi_0) \leqq F_\varepsilon(u_2, \psi_0) \quad \text{if } u_1 \leqq u_2. \tag{3.6}$$

Indeed, this follows easily from (3.1) (with $\psi_{01} = \psi_{02} = \psi_0$) and (3.2).

LEMMA 3.3. *There exists a positive constant C such that for any $u \in C^1[0, T]$, $\psi_0 \in \Psi_1$,*

$$\|F_\varepsilon(u, \psi_0) - F(u, \psi_0)\|_{L^\infty(0,T;\Psi_0)} \leqq C\sqrt{\varepsilon} + 2\|\dot{u}\|_\infty \varepsilon. \tag{3.7}$$

Proof. Denote by $\prod_\varepsilon u$ the piecewise linear function of t that coincides with $u(t_i)$ at each of the points t_i. Then

$$\begin{aligned} \|F_\varepsilon(u, \psi_0)(t) - F(u, \psi_0)(t)\| &\leqq \|F_\varepsilon(u, \psi_0)(t) - F_\varepsilon(\textstyle\prod_\varepsilon u, \psi_0)(t)\| \\ &\quad + \|F_\varepsilon(\textstyle\prod_\varepsilon u, \psi_0)(t) - F(\textstyle\prod_\varepsilon u, \psi_0)(t)\| \\ &\quad + \|F(\textstyle\prod_\varepsilon u, \psi_0)(t) - F(u, \psi_0)(t)\| \\ &\leqq 2\|u = \textstyle\prod_\varepsilon u\|_\infty + \|F_\varepsilon(\textstyle\prod_\varepsilon u, \psi_0)(t) - F(\textstyle\prod_\varepsilon u, \psi_0)(t)\| \end{aligned} \tag{3.8}$$

by (2.12) and (3.5). The last term on the right-hand side can be estimated in $[t_i, t_{i+1}]$, by

$$\begin{aligned} &\|G_\varepsilon(\textstyle\prod_\varepsilon u(t), \psi_\varepsilon(t_i)) - G(\textstyle\prod_\varepsilon u(t), \psi(t_i))\| \\ &\quad \leqq \|G_\varepsilon(\textstyle\prod_\varepsilon u(t), \psi_\varepsilon(t_i)) - G(\textstyle\prod_\varepsilon u(t), \psi_\varepsilon(t_i))\| \\ &\qquad + \|G(\textstyle\prod_\varepsilon u(t), \psi_\varepsilon(t_i)) - G(\textstyle\prod_\varepsilon u(t), \psi(t))\| \\ &\quad \leqq \varepsilon + \|\psi_\varepsilon(t_i) - \psi(t_i)\| \end{aligned}$$

by (3.2), (2.5), where

$$\psi_\varepsilon(t_i) = F_\varepsilon(\textstyle\prod_\varepsilon u, \psi_0)(t_i), \qquad \psi(t_i) = F(\textstyle\prod_\varepsilon u, \psi_0)(t_i).$$

Summing over i we conclude that the last term on the right-hand side of (3.8) is bounded by $(T/h)\varepsilon$, i.e., by $C\sqrt{\varepsilon}$. Finally, when we use the estimate

$$\|u - \textstyle\prod_\varepsilon u\|_\infty \leqq \|\dot{u}\|_\infty \varepsilon$$

in (3.8), the assertion (3.7) follows.

From the definition of F_ε in (3.4) it is clear that

$$F_\varepsilon : C[0, T] \times \Psi_0 \to \prod_i C([t_i, t_{i+1}], \Psi_0) \text{ is continuously differentiable.}$$

Now consider the state-output map E defined in (2.15). If e is not continuous, then we can regularize E by

$$E_\varepsilon(\phi, e) = E(\phi, j_\varepsilon * e). \tag{3.9}$$

The regularization W_ε of the Preisach operator W is then defined by

$$W_\varepsilon(u, e)(t) = E_\varepsilon(F_\varepsilon(u, \psi_0)(t), \tilde{T}e)$$

for the initial state $\psi_0 \equiv 0$. Since in the rest of this paper we deal exclusively with densities e that are continuous, we shall replace E_ε by E, i.e., we shall work with the simpler operators

$$W_\varepsilon(u, e)(t) = E(F_\varepsilon(u, \psi_0)(t), \tilde{T}e). \tag{3.10}$$

We introduce the mapping

$$\mathscr{E} : C([a, b]; \Psi_0) \times L^\infty(\mathbb{R}_+ \times R) \to C[a, b]$$

by

$$\mathscr{E}(\psi, e)(t) = E(\psi(t), e).$$

For $e \in C(\mathbb{R}_+ \times \mathbb{R})$ with compact support,

$$\mathscr{E}(\cdot, e) : C([a, b]; \Psi_0) \to C[a, b]$$

is continuously differentiable with derivative

$$(D_\psi \mathscr{E}(\psi, e)\chi)(t) = \int_0^\infty e(r, \psi(t)r)\chi(t)r\, dr \tag{3.11}$$

where $\psi(t)r$ is the value of the function $\psi(t)$ at the point r, and similarly for $\chi(t)r$. Indeed,

$$\left| \mathscr{E}(\psi + \chi, e)(t) - \mathscr{E}(\psi, e)(t) - \int_0^\infty e(r, \psi(t)r)\chi(t)r\, dr \right|$$

$$\leqq \int_0^\infty \left| \int_{\psi(t)r}^{\psi(t)r + \chi(t)r} |e(r, s) - e(r, \psi(t)r)|\, ds \right| dr$$

and if $\|\chi\| \to 0$ then the integrand converges to zero uniformly, and thus the integral is $o(\|\chi\|)$.

From (3.11) it follows that

$$\begin{aligned} &(D_\psi \mathscr{E}(\psi_n, e)\chi)(t) \to (D_\psi \mathscr{E}(\psi, e)\chi)(t) \text{ uniformly in } t, \\ &\qquad \text{if } \psi_n \to \psi \text{ in } C([a, b]; \Psi_0). \end{aligned} \tag{3.12}$$

Hence $W_\varepsilon(u, e)$ is continuously differentiable in u.

LEMMA 3.4. *The operator W_ε satisfies*

$$\|W_\varepsilon(u_1, e) - W_\varepsilon(u_2, e)\|_\infty \leqq \|e\|_\infty \|u_1 - u_2\|, \tag{3.13}$$

$$W_\varepsilon(u_1, e_1) \leqq W_\varepsilon(u_2, e_2) \quad \text{if } u_1 \leqq u_2,\ e_1 \leqq e_2; \tag{3.14}$$

furthermore, for any $e \in C(\mathbb{R}_+ \times \mathbb{R})$ with compact support and $N > 0$ there exists a positive constant C_0 such that

$$\|W_\varepsilon(u, e) - W(u, e)\|_\infty \leqq C_0\sqrt{\varepsilon} \quad \forall u \in W^{1,\infty}(0, T),\ \|\dot{u}\|_\infty \leqq N. \tag{3.15}$$

Proof. The estimates (3.13), (3.14) follow from (3.5), (3.6) respectively, and (3.15) follows by Lemma 3.3.

We emphasize that, in contrast to $W(u, e)$, $W_\varepsilon(u, e)$ is discontinuous as a function of time with discontinuities of order ε at the points $\{t_i\}$. (However, it is continuously differentiable in u.)

We conclude this section with some monotonicity properties with respect to translation of the measure e.

For $e \in C(\bar{\Delta})$, $\lambda \in \mathbb{R}$, set

$$(T_\lambda^1 e)(\rho_1, \rho_2) = e(\rho_1 - \lambda, \rho_2), \qquad (T_\lambda^2 e)(\rho_1, \rho_2) = e(\rho_1, \rho_2 - \lambda).$$

LEMMA 3.5. *If $e \geqq 0$, $\lambda \geqq 0$, $\phi \in \Psi_1$, then*

$$E(\phi, \tilde{T}T_\lambda^1 e) \leqq E(\phi, \tilde{T}e), \tag{3.16}$$

$$E(\phi, \tilde{T}T_\lambda^2 e) \leqq E(\phi, \tilde{T}e). \tag{3.17}$$

Proof. We have

$$\begin{aligned} E(\phi, \tilde{T}T_\lambda^1 e) &= \int_0^\infty \int_{-\infty}^\infty (\tilde{T}T_\lambda^1 e)(r, s)\, ds\, dr \\ &= \int_0^\infty \int_{-\infty}^{\phi(r)} e(s - r - \lambda, s + r)\, ds\, dr \\ &= \int_{\lambda/2}^\infty \int_{-\infty}^{\phi(r' - \lambda/2) - \lambda/2} e(s' - r', s' + r')\, ds'\, dr' \\ &\leqq \int_0^\infty \int_{-\infty}^{\phi(r')} (\tilde{T}e)(r', s')\, ds'\, dr' = E(\phi, \tilde{T}e) \end{aligned}$$

where the inequality follows from the fact that $\phi(r' - \lambda/2) \leqq \phi(r') + \lambda/2$. The proof of (3.17) is similar, provided we recall that supp $\tilde{T}e$ is contained in $\mathbb{R}_+ \times \mathbb{R}$.

4. The problems (P), (P_ε). Denote by $G(x, t; \xi, \tau)$ $(0 \leqq x, \xi \leqq a, 0 \leqq \tau < t \leqq T)$ the Green function for the heat equation in Q_T with the boundary conditions

$$G_x(0, t; \xi, \tau) = 0, \qquad G_x(a, t; \xi, \tau) + G(a, t; \xi, \tau) = 0. \tag{4.1}$$

G can be constructed by the parametrix method ([4] or [5, Chap. 1]) that involves the solution of a Volterra type integral equation; it can also be constructed by using existence and regularity theory of parabolic equations [5]. The function $G(x, t; \xi, \tau)$ satisfies in (ξ, τ) the backward heat equation and the boundary conditions (4.1), it has the same singularity as the fundamental solution

$$\frac{1}{\sqrt{4\pi(t - \tau)}} \exp\left\{-\frac{|x - \xi|^2}{4(t - \tau)}\right\},$$

and it is a positive value function. The solution of (1.1)-(1.4) can be represented in terms of G:

$$u(x,t)=\int_0^t G(x,t;a,\tau)\,W(u(0,\cdot),Bv)(\tau)\,d\tau + \int_0^t G(x,t;0,\tau)\,d\tau + \int_0^a G(x,t;\xi,0)u_0(\xi)\,d\xi. \tag{4.2}$$

THEOREM 4.1. *For any bounded measurable density function Bv there exists a unique solution of* (1.1)-(1.4).

Proof. Taking $x=0$ in (4.2) we obtain for $u(0,t)$ an integral equation of Volterra type:

$$u(0,t)=\int_0^t G(0,t;a,\tau)\,W(u(0,\cdot),Bv)(\tau)\,d\tau + \int_0^t G(0,t;0,\tau)k(\tau)\,d\tau + \int_0^a G(0,t;\xi,0)u_0(\xi)\,d\xi. \tag{4.3}$$

The integrand in the first integral is a nonanticipative Lipschitz continuous functional in $u(0,\cdot)$. By a standard method it follows that this integral equation has a unique solution $u(0,t)$, and then a solution $u(x,t)$ is obtained by solving (1.1)-(1.4) (since the right-hand side of (1.4) is now well defined). Uniqueness follows from the uniqueness of the solution $u(0,\cdot)$ of the integral equation (4.3).

We now specify the choice of admissible densities $e=Bv$ in problem (P). Let $\sigma(\rho)=\sigma(\rho_1,\rho_2)$ be a specific spherically symmetric continuously differentiable function such that

$$\sigma(\rho)=0 \quad \text{if } |\rho|\geqq 1, \qquad \sigma(\rho)>0 \quad \text{if } |\rho|<1,$$
$$\iint \sigma(\rho_1,\rho_2)\,d\rho_1\,d\rho_2=1, \tag{4.4}$$
$$\frac{\partial\sigma}{\partial\rho_i}>0\ (<0) \quad \text{if } \rho_i>0\ (<0), \quad |\rho|<1, \quad i=1,2,$$

and set

$$\sigma_\eta(\rho)=\frac{1}{\eta^2}\sigma\left(\frac{\rho_1}{\eta},\frac{\rho_2}{\eta}\right), \qquad \eta>0. \tag{4.5}$$

We define the control set $K\subset V$ for problem (P) setting

$$V=\mathbb{R}^3,$$
$$K=\{v=(\eta,\alpha,\beta):\ \eta_0\leqq\eta\leqq 1,\ \eta\leqq\alpha,\ \alpha+\eta\leqq\beta\leqq 1-\eta\} \tag{4.6}$$

where $\eta_0>0$ is a given number, $\eta_0\ll 1$. We define $B:\mathbb{R}_+\times\mathbb{R}^2\to L^\infty(\Delta)$ by

$$(Bv)(\rho_1,\rho_2)=\sigma_\eta(\rho_1-\alpha,\rho_2-\beta).$$

Then the admissible densities $e=Bv$ form a three parameter family, approximating Dirac measures that represent the thermostats.

THEOREM 4.2. *There exists an optimal control $v_*\in K$ of problem (P), that is,*

$$J(v_*)=\min_{v\in K} J(v).$$

Proof. Let v_n be a minimizing sequence and set $u_n = u(v_n)$. By parabolic estimates

$$\|u_n(0, \cdot)\|_{C^1[0,T]} \leqq C_0.$$

Using the representation (4.2) it is easy to see that a subsequence is convergent to a solution of problem (P).

Remark 4.1. Theorem 4.2 can be extended to the case where v varies in an infinite-dimensional space; the proof is similar to that of Theorem 3.1 in [5].

We now turn to the approximating problems (P_ε) whereby W in (1.4) is replaced by W_ε, and $J(v)$ is replaced by

$$J_\varepsilon(v) = \int_0^T \Phi(u_\varepsilon(0, t), t)\, dt + h(v) + |v - v_*|^2; \tag{4.7}$$

here $u_\varepsilon = u_\varepsilon(v)$ is the solution of (1.1)-(1.4) with W replaced by W_ε; the existence and uniqueness of $u_\varepsilon(v)$ is proved as in Theorem 4.1.

Problem (P_ε). Find v_ε in K such that

$$J_\varepsilon(v_\varepsilon) = \min_{v \in K} J_\varepsilon(v).$$

The proof of Theorem 4.1 applies also to problem (P_ε), showing that a solution v_ε exists.

THEOREM 4.3. *As* $\varepsilon \to 0$, $|v_\varepsilon - v_*| \to 0$ *and*

$$\|u_\varepsilon(v_\varepsilon) - u_*\|_{L^\infty(Q_T)} \to 0.$$

Proof. For convenience we write $J(v)$ also as $J(v, u(v))$ if $u(v)$ is the solution of (1.1)-(1.4) and as $J(v, u_\varepsilon(v))$ if $u_\varepsilon(v)$ is the solution of (1.1)-(1.4) with W replaced by W_ε, and set

$$J_\varepsilon(v, u) = J(v, u) + |v - v_*|^2.$$

Then, by the optimality of v_ε,

$$\begin{aligned} J_\varepsilon(v_\varepsilon, u_\varepsilon(v_\varepsilon)) &\leqq J_\varepsilon(v_*, u_\varepsilon(v_*)) \\ &= J(v_*, u_*) + R_\varepsilon \end{aligned} \tag{4.8}$$

where

$$R_\varepsilon = J(v_*, u_\varepsilon(v_*)) - J(v_*, u_*).$$

Analogously to (4.3),

$$\begin{aligned} u_\varepsilon(0, t) = &\int_0^t G(0, t; a, \tau)\, W_\varepsilon(u_\varepsilon(0, \cdot)Bv)(\tau)\, d\tau \\ &+ \int_0^t G(0, t; 0, \tau) k(\tau)\, d\tau + \int_0^a G(0, t; \xi, 0) u_0(\xi)\, d\xi. \end{aligned} \tag{4.9}$$

Subtracting this from (4.3) and using (3.15), we get

$$\|u_\varepsilon(v)(0, \cdot) - u(v)(0, \cdot)\|_\infty \leqq C\sqrt{\varepsilon} + \int_0^t |W(u_\varepsilon(0, \cdot), Bv) - W(u(0, \cdot), Bv)|\, d\tau.$$

Using (2.20) and Gronwall's inequality we deduce that

$$\|u_\varepsilon(v)(0, \cdot) - u(v)(0, \cdot)\|_\infty \to 0$$

as $\varepsilon \to 0$. It follows that $R_\varepsilon \to 0$, and (4.8) yields

$$\lim_{\varepsilon \to 0} J(v_\varepsilon, u_\varepsilon(v_\varepsilon)) + \overline{\lim_{\varepsilon \to 0}} |v_\varepsilon - v_*|^2 \leqq J(v_*, u_*). \tag{4.10}$$

On the other hand, for a subsequence,

$$v_\varepsilon \to \hat{v} \quad \text{and} \quad u_\varepsilon(v_\varepsilon) \to \hat{u} \quad \text{in } C(Q_T);$$

comparing (4.9) with (4.3) when $v = \hat{v}$, $u = u(\hat{v})$, and applying (3.15), we find that $u(\hat{v}) = \hat{u}$. It follows that

$$J(v_\varepsilon, u_\varepsilon(v_\varepsilon)) \to J(\hat{v}, u(\hat{v})) \geqq J(v_*, u_*)$$

by the optimality of v_*. Substituting this into (4.10) we find that $|v_\varepsilon - v_*| \to 0$, and the theorem follows.

5. Optimality conditions for (P_ε). Since W_ε is differentiable, the optimality conditions (1.12) can be rigorously derived for any solution $(v_\varepsilon, u(v_\varepsilon))$ of problem (P_ε), with a minor change due to the term $|v - v_*|^2$ in J_ε; it will suffice to write the conclusion.

We begin by setting

$$D_u W_\varepsilon = D_u W_\varepsilon(u_\varepsilon, Bv_\varepsilon), \qquad u_\varepsilon = u_\varepsilon(v_\varepsilon)(0, \cdot), \tag{5.1}$$

$$D_e W_\varepsilon = D_\varepsilon W_\varepsilon(u_\varepsilon, Bv_\varepsilon) DB(v_\varepsilon) \tag{5.2}$$

and introducing $D_u W_\varepsilon^*$ by

$$\int_0^T p(t)(D_u W_\varepsilon z)(t)\, dt = \int_0^T z(t)(D_u W_\varepsilon^* p)(t)\, dt \quad \forall p, z \in C[0, T]. \tag{5.3}$$

LEMMA 5.1. *For any $n \in \mathbb{N}$, set $h = 1/n$ and $\varepsilon = h^2$. Then there exist nonnegative functions $w_\varepsilon(t)$, $w_i^\varepsilon(t)$, continuous on every $[t_j, t_{j+1}]$ where $t_j = jh$, such that, for every $p \in L^2[0, T]$,*

$$D_u W_\varepsilon^* p(t) = w_\varepsilon(t) p(t) + \sum_i \left\{ \int_t^T p(s) w_i^\varepsilon(s)\, ds \right\} \delta_{t_i}(t) \tag{5.4}$$

where $\delta_{t_i}(t)$ is the Dirac measure at $t = t_i$.

Proof. From (3.4) we get, for $t_i \leqq t \leqq t_{i+1}$,

$$\begin{aligned}(D_u F_\varepsilon(u, \psi_0)z)(t) &= D_x G_\varepsilon(u(t), F_\varepsilon(u, \psi_0)(t_i - 0))z(t) \\ &\quad + [D_\psi G_\varepsilon(u(t), F_\varepsilon(u, \psi_0)(t_i - 0))](D_u F_\varepsilon(u, \psi_0)z)(t_i - 0).\end{aligned}$$

Defining the linear continuous operators $A(t)$, $A_j : \Psi_0 \to \Psi_0$ and $B(t)$, $B_j : \mathbb{R} \to \Psi_0$ by

$$\begin{aligned}A(t) &= D_\psi G_\varepsilon(u(t), F_\varepsilon(u, \psi_0)(t_i - 0)), \\ B(t) &= D_x G_\varepsilon(u(t), F_\varepsilon(u, \psi_0)(t_i - 0)), \\ A_{i+1} &= A(t_{i+1} - 0), \qquad B_{i+1} = B(t_{i+1} - 0),\end{aligned}$$

we get, taking $t = t_{i+1}$,

$$(D_u F_\varepsilon(u, \psi_0)z)(t_{i+1} - 0) = B(t_{i+1})z(t_{i+1}) + A_{i+1}(D_u F_\varepsilon(u, \psi_0)z)(t_i - 0).$$

It follows that for $t_i \leqq t \leqq t_{i+1}$

$$\begin{aligned}(D_u F_\varepsilon(u, \psi_0)z)(t) &= B(t)z(t) + A(t)B_i z(t_i) + A(t)A_{i-1}B_{i-1}z(t_{i-1}) \\ &\quad + \cdots + A(t)A_{i-1} \cdots A_1 B_1 z(t_1).\end{aligned} \tag{5.5}$$

Note that by (2.9)

$$A(t) \geqq 0, \qquad B(t) \geqq 0. \tag{5.6}$$

Now since

$$D_u W_\varepsilon : C[0, T] \to \prod_i C[t_i, t_{i+1}],$$

we have

$$D_u W_\varepsilon^* : \left(\prod_i C[t_i, t_{i+1}]\right)^* \to C[0, T]^*;$$

in particular, $D_u W_\varepsilon$ maps $C[0, T]$ into $L^2[0, T]$ and $D_u W_\varepsilon^*$ maps $L^2[0, T]$ into $C[0, T]^*$. Next, by (3.10), (5.5), with $\psi_\varepsilon = f_\varepsilon(u_\varepsilon, \psi_0)$,

$$\begin{aligned}(D_u W_\varepsilon z)(t) &= \int_0^\infty e(r, \psi_\varepsilon(t)r)((D_u F_\varepsilon \cdot z)(t))(r)\, dr \\ &= \int_0^\infty e(r, \psi_\varepsilon(t)r)[B(t)z(t) + A(t)B_i z(t_i) + \cdots](r)\, dr\end{aligned}$$

so that

$$(D_u W_\varepsilon(u_\varepsilon, \psi_\varepsilon)z)(t) = w_\varepsilon(t)z(t) + \sum_{0<t_i<t} w_i^\varepsilon(t) z(t_i)$$

where $w_\varepsilon(t) \geqq 0$, $w_i^\varepsilon(t) \geqq 0$ by (5.6) and w_ε, w_i^ε are continuous on $[t_j, t_{j+1}]$. Finally,

$$\begin{aligned}\int_0^T p(t)(D_u W_\varepsilon(u_\varepsilon, \psi_0)z)(t)\, dt &= \int_0^T \left[p(t)w_\varepsilon(t)z(t) + p(t) \sum_{0<t_i<t} w_i^\varepsilon(t) z(t_i) \right] dt \\ &= \int_0^T p(t)w_\varepsilon(t)z(t)\, dt + \sum_i \left[\int_{t_i}^T p(t) w_i^\varepsilon(t)\, dt \cdot z(t_i) \right],\end{aligned}$$

which is precisely the assertion (5.4).

Denote by $p = p_\varepsilon$ the solution of

$$\begin{aligned}&p_t + p_{xx} = 0 \quad \text{in } Q_T, \\ &p(x, T) = 0, \qquad 0 < x < a, \\ &-p_x(0, t) = (D_u W_\varepsilon^* p(a, \cdot))(t) + \Phi_u(u_\varepsilon(0, t), t), \qquad 0 < t < T, \\ &p_x(a, t) + p(a, t) = 0, \qquad 0 < t < T.\end{aligned} \tag{5.7}$$

Representing $p_\varepsilon(x, t)$ by means of Green's function G (cf. (4.2)) we obtain for $p(a, t)$ an integral equation of Volterra type (cf. (4.3)). Its solution yields a unique solution for (5.7).

The optimality conditions for $(v_\varepsilon, u_\varepsilon)$, whose derivation is the same as for (1.12), are:

$$\pm\left\{ \int_0^T p_\varepsilon(a, t)(D_e W_\varepsilon \cdot w)(t) + Dh(v_\varepsilon)w + 2(v_\varepsilon - v_*) \cdot w \right\} \geqq 0 \quad \forall w \in K^{\pm}(v_\varepsilon). \tag{5.8}$$

To derive properties for v_ε (and v_*) from (5.8), we need to establish some auxiliary properties for $D_e W_\varepsilon$, $D_u W_\varepsilon^*$ and p_ε; this is done in the next section.

6. Auxiliary properties of DW_ε.

LEMMA 6.1. *For any $p \in C[0, T]$,*

$$(D_u W_\varepsilon^*(u, e)p(\cdot))(t) \geqq 0 \quad \text{if } p \geqq 0. \tag{6.1}$$

Indeed, this follows from Lemma 5.1.

LEMMA 6.2. *If* $\Phi_u(u, s) > 0$, *then*

$$p_\varepsilon(x, t) > 0 \quad \text{for } 0 \leqq x \leqq a, \quad 0 \leqq t < T. \tag{6.2}$$

Proof. We can recast the proof of existence of the solution $p = p_\varepsilon$ in the following way. Given $p_0(t) \geqq 0$ solve (5.7) with $p(a, t) = p_0(t)$ and denote the solution by $\tilde{p}(x, t)$. Since $D_u W_\varepsilon^* p_0 \geqq 0$ (by Lemma 6.1), the maximum principle gives $\tilde{p}(x, t) \geqq 0$ if $0 \leqq x \leqq a$, $0 < t < T$ and, in particular, $p_1(t) \equiv \tilde{p}(a, t) \geqq 0$. Consider the mapping $S: p_1 = Sp_0$. Representing S (by means of Green's function G) as an integral operator of Volterra type, we deduce that S must have a fixed point $\tilde{p}(t)$, which then corresponds to a solution $p(x, t)$ of (5.7) with $p(a, t) = \tilde{p}(t) \geqq 0$. The conclusion (6.2) now readily follows.

LEMMA 6.3. *Write* $v_\varepsilon = (\eta_\varepsilon, \alpha_\varepsilon, \beta_\varepsilon)$, $w_1 = (0, 1, 0)$, $w_2 = (0, 0, 1)$. *Then*

$$D_e W_\varepsilon(u_\varepsilon, Bv_\varepsilon) DBv_\varepsilon \cdot w_i = W_\varepsilon(u_\varepsilon, e_i') \tag{6.3}$$

where $u_\varepsilon = u_\varepsilon(0, \cdot)$,

$$\begin{gathered} e_i' = -\frac{1}{\eta^3} D_i \sigma\left(\frac{\rho_1 - \alpha}{\eta}, \frac{\rho_2 - \beta}{\eta}\right), \\ D_1\sigma(s_1, s_2) = \frac{\partial}{\partial s_1}\sigma(s_1, s_2), \qquad D_2\sigma(s_1, s_2) = \frac{\partial}{\partial s_2}\sigma(s_1, s_2). \end{gathered} \tag{6.4}$$

The proof follows directly from the definition of $D_e W_\varepsilon$.

LEMMA 6.4. $W(u_\varepsilon, e_i') \leqq 0$, $W_\varepsilon(u_\varepsilon, e_i') \leqq 0$.

Indeed, this follows from Lemma 3.5.

LEMMA 6.5. *For any* $\delta > 0$, $N > 0$, *there exists a positive constant* $c_0 = c_0(\delta, N)$ *such that for any* $(\eta, \alpha, \beta) \in K$ (K *as in* (4.6)) *and* $u \in C^1[0, T]$ *with* $0 \leqq u(0) \leqq \beta - \eta$, $\|\dot{u}\|_\infty \leqq N$, *either*

$$W(u, e)(t) \leqq \delta \quad \forall 0 \leqq t \leqq T, \tag{6.5}$$

or

$$\int_0^{T-\delta} W(u, e_2')(t)\, dt \leqq -c_0, \tag{6.6}$$

where

$$e(\rho_1, \rho_2) = \frac{1}{\eta^2}\sigma\left(\frac{\rho_1 - \alpha}{\eta}, \frac{\rho_2 - \beta}{\eta}\right) \tag{6.7}$$

and e_2' *is defined by* (6.4).

Proof. We may assume that $\delta < \frac{1}{2}$. Suppose (6.5) is not true. Since the initial state is $\psi_0 \equiv 0$ so that $W(u, e)(0) = 0$, there must exist a first time $t = t_*$ such that $W(u, e)(t_*) \geqq \delta$. The internal state $\psi(t_*)$ (at $t = t_*$) in the Preisach plane (ρ_1, ρ_2) is a horizontal segment; see Fig. 3. Since

$$\begin{aligned} W(u, e)(t_*) &= \frac{1}{\eta^2}\int_{A^-(t_*)} \sigma\left(\frac{\rho_1 - \alpha}{\eta}, \frac{\rho_2 - \beta}{\eta}\right) d\rho_1\, d\rho_2 \\ &= \int_{A_0^-} \sigma(\rho_1 - \alpha, \rho_2 - \beta)\, d\rho_1\, d\rho_2 = \delta \qquad \left(A_0^\pm = \frac{1}{\eta} A^\pm(t_*)\right), \end{aligned}$$

it follows that

$$\text{meas } A_0^+ = \theta \text{ meas } A_0^- \quad \text{where } \theta = \theta(\delta) \in (0, 1); \tag{6.8}$$

θ is independent of η, α, β.

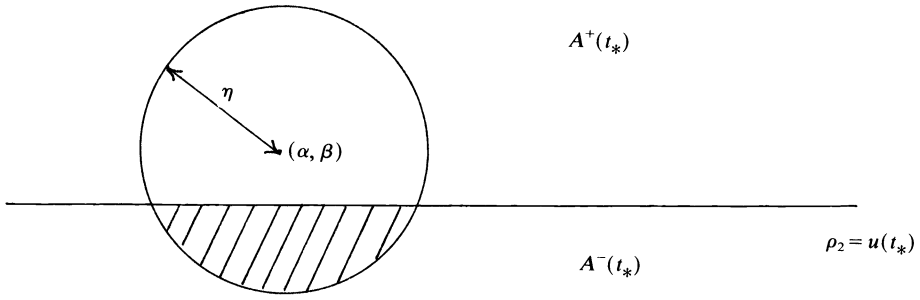

FIG. 3

Recalling that

$$\frac{\partial}{\partial \rho_2}\sigma(\rho_1, \rho_2) < 0 \quad (>0) \quad \text{if } \rho_1^2 + \rho_2^2 < 1 \quad \text{and } \rho_2 > 0 \quad (\rho_2 < 0), \tag{6.9}$$

and using (6.8), we can then estimate

$$\begin{aligned} W(u, e_2')(t_*) &= \int_{A^-(t_*)} e_2'(\rho_1, \rho_2)\, d\rho_1\, d\rho_2 \\ &= -\frac{1}{\eta}\int_{A_0^-} D_2\sigma(\rho_1 - \alpha, \rho_2 - \beta)\, d\rho_1\, d\rho_2 \leqq -\frac{c}{\eta} \qquad (c>0), \end{aligned} \tag{6.10}$$

where c is independent of η, α, β.

From (2.21) we see that the internal states $\psi(t)$ and $\psi(t_*)$ satisfy

$$\|\psi(t) - \psi(t_*)\| \leqq C|t - t_*|.$$

Therefore

$$|W(u, e_2')(t) - W(u, e_2')(t_*)| \leqq \int_{A_\eta} |\tilde{T} e_2'| \tag{6.11}$$

where A_η is the intersection of a disc with radius η with a polygonal strip of width $C|t - t_*|$ in the (r, s)-plane. Substituting $r \to r\eta$, $s \to s\eta$ we can estimate the right-hand side of (6.11) by $C|t - t_*|/\eta^2$, and then, by (6.10),

$$W(u, (\cdot), e_2')(t) < -\frac{c}{2\eta} \quad \text{if } |t - t_*| \leqq \frac{c}{2C}\,\eta. \tag{6.12}$$

The assertion (6.6) now follows from Lemma 6.4 and (6.12).

We finally state an extension of Lemma 6.5 to W_ε.

LEMMA 6.6. *For any $\delta > 0$, $N > 0$ there exist $c_0 = c_0(\delta, N) > 0$ and $\varepsilon_0 = \varepsilon(\delta, N, \eta) > 0$ such that for any $(\eta, \alpha, \beta) \in K$, $u \in C^1[0, T]$ with $0 \leqq u(0) \leqq \beta - \eta$, $\|\dot{u}\|_\infty \leqq N$, either*

$$W_\varepsilon(u(\cdot), e)(t) \leqq \delta \quad \forall 0 \leqq t \leqq T, \tag{6.13}$$

or

$$\int_0^{T-\delta} W_\varepsilon(u(\cdot), e_2')(t)\, dt \leqq -c_0, \tag{6.14}$$

provided $\varepsilon < \varepsilon_0$.

Proof. We apply Lemma 6.5 with δ replaced by $\delta/2$ and then use (3.15) to deduce that if $C_0\sqrt{\varepsilon} < \delta/2$ then either (6.13) holds or else (6.6) holds. Next, by (3.15) and (3.7),

$$\int_0^T |W_\varepsilon(u(\cdot), e_2') - W(u(\cdot), e_2')| \leqq \frac{C}{\eta^2} \|F_\varepsilon(u, \psi_0) - F(u, \psi_0)\|_\infty \leqq \frac{C_1\sqrt{\varepsilon}}{\eta^2}.$$

From this inequality and (6.6) we deduce the inequality (6.14).

7. Properties of the optimal control. The results of §§ 5 and 6 can be used to derive properties for the optimal control v_*. For simplicity we illustrate this in the case where η and α are fixed, whereas β is the only control variable, and $-\alpha + \eta \leqq \beta \leqq 1 - \eta$. We assume that

$$J(v) = \int_0^T \Phi(u(0, t), t)\, dt + \mu h_0(\beta) \tag{7.1}$$

where

$$\begin{aligned} &\Phi_u(u, g(t)) \geqq \gamma > 0 \quad \text{if } u \geqq 0, \quad 0 \leqq t \leqq T, \\ &h_0'(\beta) > 0 \quad \text{if } 0 \leqq \beta \leqq 1 \end{aligned} \tag{7.2}$$

and μ is a positive constant. As before we fix $\psi_0 \equiv 0$. We also assume that $u(0, 0) \leqq \alpha$.

Denote by β_*, u_* the solution to problem (P) with J given by (7.1), and by β_ε, u_ε the corresponding solution of problem (P_ε). The optimality conditions (5.8) become

$$\int_0^T p_\varepsilon(a, t)\, W_\varepsilon(u_\varepsilon(0, \cdot), e_1')(t)\, dt + \mu h_0'(\beta_\varepsilon) + 2(\beta_\varepsilon - \beta_*) \geqq 0 \quad \text{if } \beta_\varepsilon < 1 - \eta. \tag{7.3}$$

THEOREM 7.1. *For any small δ there exists a (small) $\mu_0 > 0$ depending only on δ, γ, a, T and $\|h_0'\|_\infty$ such that if $\mu < \mu_0$ then*

$$W(u_*(0, \cdot), e_*)(t) < \delta \quad \forall 0 \leqq t \leqq T, \tag{7.4}$$

provided $\beta_* < 1 - \eta$.

Taking $\Phi(u, g) = |u - g(t)|^2$, Theorem 7.1 can be interpreted as follows. If the desired temperature $g(t)$ is lower than any achievable temperature u and if the cost μh_0 of cooling is small ($\mu < \mu_0$) then, in the optimal situation, the thermostat is either set at $\beta = 1 - \eta$ so as to cool the rod at $x = a$ as much as possible, or it is set at such value of β so that essentially no heating will occur at $x = a$ (i.e., (7.4) holds).

Proof. Representing $p_\varepsilon(x, t)$ by Green's function and using the estimate $\Phi_u \geqq \gamma > 0$ and Lemmas 6.1, 6.2, we find that

$$p_\varepsilon(a, t) \geqq \int_0^t G(a, t; 0, \tau) \Phi_u(u_\varepsilon(0, \tau), \tau)\, d\tau \geqq c_1 t,$$

with $c_1 > 0$. If (6.14) holds, then upon using Lemma 6.4 we derive the inequality

$$\int_0^T W_\varepsilon(u_\varepsilon(0,), e_2')(t) p_\varepsilon(a, t)\, dt \leqq -c_2 < 0,$$

which is a contradiction to (7.3) if $\beta_\varepsilon < 1 - \eta$ and $\mu < \mu_0$, μ_0 small. Thus by Lemma 6.6 (here we use the assumptions that $u(0, 0) \leqq \alpha \leqq \beta - \eta$) it follows that (6.13) must be satisfied if $\beta_\varepsilon < 1 - \eta$ and, consequently, (7.4) must hold if $\beta_* < 1 - \eta$.

Remark 7.1. As η decreases (7.4) continues to hold provided $\beta_* = \beta_*(\eta) < 1 - \eta$. However, we cannot deal directly with the case $\eta = 0$ (i.e., with the case where the Preisach measure is a Dirac measure).

8. Generalizations. The results of §§ 4–7 extend to other functionals such as

$$J(v)=\int\int_{Q_T}\Phi(u(x,t),x,t)\,dx\,dt+h(v) \tag{8.1}$$

or

$$J(v)=\int_0^a \Phi(u(x,T),x)\,dx+h(v). \tag{8.2}$$

Thus in case (8.1) we formally replace (1.10) by

$$\begin{aligned} &p_t+p_{xx}=\Phi_u(u_*(x,t),x,t) \quad \text{in } Q_T,\\ &p(x,T)=0, \qquad 0<x<a,\\ &-p_x(x,0)=[D_uW^*p(a,\cdot)](t), \qquad 0<t<T,\\ &p_x(a,t)+p(a,t)=0, \qquad 0<t<T \end{aligned} \tag{8.3}$$

and $p(x,t)>0$ if $\Phi_u<0$; the optimality conditions (1.12) remain valid and the assertion of Theorem 7.1 continues to hold.

The results of §§ 4–7 extend to models whereby (1.4) is replaced by

$$u_x(a,t)+u(a,t)=m(t)+W(u(x_0,\cdot),Bv)(t) \tag{8.4}$$

with $m(t)$ a given function and $x_0=0$ or $x_0=a$. The case $x_0=a$ may be interpreted as a generalized cooling law with hysteresis.

In the above problems we may also take $Bv\equiv e$ fixed and consider the function $k(t)$ (in (1.3)) as the control variable, say in the class

$$K=\left\{0\leqq k(t)\leqq 1,\ \int_0^T k(t)\,dt=N\right\},$$

where $T>N$. Assuming $\Phi_u<0$ we find, for the solution p_ε of the corresponding adjoint problem of (P_ε), that

$$\int_0^T p_\varepsilon l\leqq 2\int_0^T (k_\varepsilon-k_*)l \quad \text{if } k_*+\lambda l\in K \quad \forall\lambda>0,\ \lambda \text{ small}.$$

Since $\int|k_\varepsilon-k_*|^2\to 0$ as $\varepsilon\to 0$, it follows that except for a subset of $[0,T]$ of measure $\leqq\eta_\varepsilon$

$$\begin{aligned} &k_\varepsilon=1 \quad \text{on } \{p_\varepsilon>\gamma_\varepsilon+\eta_\varepsilon\},\\ &k_\varepsilon=0 \quad \text{on } \{p_\varepsilon<\gamma_\varepsilon-\eta_\varepsilon\} \end{aligned} \tag{8.5}$$

for some number γ_ε, and $\eta_\varepsilon\to 0$ if $\varepsilon\to 0$. If we can show that the sets $\{a<p_\varepsilon<b\}$ have uniformly small measures as $b-a$ becomes small, then (8.5) would yield a bang-bang principle.

Consider finally a model where the last condition in (0.1) is replaced not by (1.4) but by

$$u_x(a,t)+u(a,t)=1-W(u(0,\cdot),Bv)(t) \tag{8.6}$$

(W appears here with negative coefficient). This represents the standard thermostat control. For this case we can still solve the adjoint problem for p_ε; however, the boundary condition

$$p_x(0,t)=W_u^*(u(0,\cdot),e)p(a,t)+\Phi_u(u(0,t),t) \tag{8.7}$$

is such that we cannot establish that p_ε has a fixed sign, no matter what Φ_u is. Therefore we are unable to derive specific properties of the optimal control for this model.

REFERENCES

[1] M. BROKATE, *Optimale Steuerung von gewöhnlichen Differentialgleichungen mit Nichtlinearitäten vom Hysteresis-Typ*, Peter Lang, Frankfurt, 1987.

[2] ———, *Optimal control of* ODE *systems with hysteresis nonlinearities*, in Trends in Mathematical Optimization, Birkhäuser, Basel, 1988.

[3] M. BROKATE AND A. VISINTIN, *Properties of the Preisach model for hysteresis*, submitted.

[4] S. D. EIDELMAN, *Parabolic Systems*, North-Holland, Amsterdam, 1969.

[5] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964, reprinted, Krieger, Melbourne, FL.

[6] A. FRIEDMAN AND K.-H. HOFFMANN, *Control of free boundary problems with hysteresis*, SIAM J. Control Optim., 26 (1988), pp. 42–55.

[7] M. A. KRASNOSELSKII AND A. V. POKROVSKII, *Modeling transducers with hysteresis by means of continuous systems of relays*, Soviet Math., 17 (1976), pp. 447–451.

[8] ———, *Systems with Hysteresis*, Nauka, Moscow, 1983. (In Russian.)

[9] F. PREISACH, *Über die magnetische Nachwirkung*, Z. Phys., 94 (1935), pp. 277–302.

[10] A. VISINTIN, *On the Preisach model for hysteresis*, Nonlinear Anal., 8 (1984), pp. 977–996.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 718-736, July 1989

003

CONVERGENCE OF FINITE ELEMENT APPROXIMATIONS TO STATE CONSTRAINED CONVEX PARABOLIC BOUNDARY CONTROL PROBLEMS*

WALTER ALT† AND UWE MACKENROTH†

Abstract. This paper is concerned with the numerical solution of state constrained parabolic boundary control problems by finite element approximations. Error estimates for the optimal values, and in the coercive case for the optimal solutions, are derived. These estimates are used to prove convergence results under rather weak assumptions. In the noncoercive case a bang-bang principle is shown to obtain convergence results for the discrete controls also. A discussion of several numerical examples concludes the paper.

Key words. optimal control, state constraints, numerical solution, finite element method

AMS(MOS) subject classifications. 49D15, 65K10

1. Introduction. In this paper we are concerned with optimal control problems governed by a parabolic differential equation of the following kind:

$$\frac{\partial y}{\partial t} + Ay = 0, \tag{1.1}$$

$$\alpha y|_\Sigma + \beta \frac{\partial y}{\partial n_A} = u, \tag{1.2}$$

$$y(0) = 0. \tag{1.3}$$

Here, A denotes the second order elliptic operator

$$A := -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) + a_0$$

with smooth coefficients a_{ij}, a_0, which are assumed to be defined on the closure of a bounded open domain $\Omega \subset \mathbb{R}^n$ with smooth boundary Γ. In addition we suppose that

$$a_{ij} = a_{ji}, \qquad i,j = 1, \cdots, n,$$

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq c \sum_{i=1}^{n} \xi_i^2 \qquad \forall x \in \bar{\Omega}, \quad \forall \xi \in \mathbb{R}^n.$$

Let $\alpha \geqq 0$, $\beta \geqq 0$ be such that $\alpha + \beta \neq 0$ and define $\Sigma :=]0, T] \times \Gamma$ with fixed $T > 0$. It is well known that under these conditions, (1.1)-(1.3) have for each $u \in L^\infty(\Sigma)$ a unique solution $y = y(u)$ in $C([0, T]; L^2(\Omega))$ depending continuously on u.

Moreover, let $\nu \geqq 0$, $\rho_1 < \rho_2$, $y_T \in L^2(\Omega)$, and for all $t \in [0, T]$, let a closed convex set $C(t) \subset L^2(\Omega)$ be given. Independent of the considered space, we denote the L^2-norm on this space by $\|\cdot\|$. We are now in a position to introduce the following parabolic optimal control problem.

(P) Minimize $\|y(T) - y_T\|^2 + \nu\|u\|^2$

subject to

$u \in L^\infty(\Sigma)$, $y \in C([0, T]; L^2\Omega))$ such that (1.1)-(1.3) hold and

$\rho_1 \leqq u(t) \leqq \rho_2 \quad \forall' t \in [0, T]$,

$y(t) \in C(t) \qquad \forall t \in [0, T]$.

* Received by the editors April 6, 1987; accepted for publication (in revised form) November 21, 1988.
† Mathematisches Institut, Universität Bayreuth, Postfach 101251, D-8580 Bayreuth, West Germany.

A typical example for $C(t)$ is given by

$$C(t) = \{v \in L^2(\Omega) \,|\, (v, g_\nu) \leqq \alpha_\nu(t),\ \nu = 1, \cdots, l\}$$

with $g_1, \cdots, g_l \in L^2(\Omega)$ and real-valued functions α_ν ($(\cdot, \cdot)$ denotes the scalar product of an L^2-space).

The main feature of problem (P) is the presence of a state constraint. There is already a rather long list of investigations on parabolic optimal control problems with a state constraint. We mention here only Barbu and Precupanu [2], Lasiecka [6], Mackenroth [9], [10], [12], and Tröltzsch [18]. These papers are mainly concerned with theoretical aspects such as characterization and qualitative behaviour of optimal controls. In contrast to this, we are concerned with numerical approximations of the optimal values and the optimal solutions.

Parabolic optimal control problems without state constraints are analyzed with respect to numerical solutions in great detail. In the coercive case ($\nu > 0$), the optimal control is in a certain sense regular, and therefore it is possible to derive rates of convergence for numerical approximations (cf. Lasiecka [7], Malanowski [14]). However, no general regularity conditions are known for the bang-bang case ($\nu = 0$), and as a consequence, only convergence in the L^p-norm without convergence rates can be shown (cf. Lasiecka [8], Knowles [5]). Thus, since for the state constrained problem (P) no regularity results are available, we cannot expect to get more than mere convergence. For different, but in some sense related, aspects of the theory, compare also Fattorini [4] and Tröltzsch [19].

The plan of the paper is as follows. In § 2 we put the problem into a more general framework and introduce discretizations in a rather abstract setting. Without requiring additional effort, this leads to a more transparent argumentation and, of course, to more generality; the general results could, for instance, also be applied to hyperbolic control problems. The main result of this section is Theorem 2, where error estimates for the optimal values and the optimal controls are given.

In § 3 we introduce a semidiscretization of (P). More precisely, only the space variable will be discretized while the time variable remains continuous. The finite element method is used to approximate both the controls and the states. An application of Theorem 2 then shows that under reasonable assumptions we obtain convergence for the optimal values and also in the coercive case for the optimal solutions. It is quite obvious that these results depend on the convergence properties of the finite element method for the partial differential equation (1.1)-(1.3). In this context we shall make use of convergence results given by Lasiecka [8] for Dirichlet boundary conditions and by Knowles [5] for Neumann and mixed boundary conditions.

In the noncoercive case, the general theory of § 2 only gives convergence of the optimal values. In § 4 we therefore use a more refined analysis to prove a bang-bang principle for the state constraint problem (P). Then it can be shown that, roughly speaking, the discrete optimal controls converge to the optimal control u_0 of (P) on the set $M \times \Gamma$, where M is the set of all $t \in [0, T]$ where the state constraint is not active.

The last section is entirely devoted to the discussion of numerical examples. The numerical results show that the described method leads within a reasonable computing time to a good approximation of the optimal value and the optimal control.

2. Error estimates for discretizations of an abstract control problem. Let U, Z be Banach spaces, E, W Hilbert spaces, $U_{ad} \subset U$, $K \subset Z$ closed convex sets, and let S be a linear continuous operator from U into $Z \times E$ (i.e., $S \in \mathscr{L}(U, Z \times E)$). Moreover, suppose that $U \subset W$ holds with continuous injection and assume that U is the dual

space of a separable Banach space $\tilde{U}$ such that $S^*(Z^* \times E^*) \subset \tilde{U}$. Then, with $\nu \geqq 0$, $y_T \in E$ we consider the following abstract optimal control problem.

($\hat{\mathrm{P}}$) Minimize $\| p_2 S_u - y_T \|_E^2 + \nu \| u \|_W^2$

subject to

$u \in U_{ad}, \qquad p_1 S_u \in K.$

By p_1 (respectively, p_2) we denote the canonical projection of $Z \times E$ onto Z (respectively, E). The proof of the following lemma is standard.

LEMMA 2.1. *Suppose U_{ad} is bounded or $\nu > 0$, $U = W$, and the feasible set is nonempty. Then* (P) *has an optimal solution u_0 which is unique for $\nu > 0$.*

The Slater condition will be needed for the optimality conditions as well as for the error estimates.

(SL) There is an $\bar{u} \in U_{ad}$ such that $p_1 S\bar{u} \in \operatorname{int} K$.

For $x \in X$, $\lambda \in X^*$ (X a Banach space) let $\langle x, \lambda \rangle := \lambda(x)$ and

$$f(u) := \| p_2 Su - y_T \|_E^2 + \nu \| u \|_W^2 \qquad \forall u \in U.$$

The next lemma is also well known.

LEMMA 2.2. *Suppose that* (SL) *is fulfilled. Then $u_0 \in U$ is optimal for* (P) *if and only if u_0 is feasible and there is a $\zeta \in Z^*$ such that*

$$\langle u - u_0, f'(u_0) + (p_1 S)^* \zeta \rangle \geqq 0 \qquad \forall u \in U_{ad}, \tag{2.1}$$

$$\langle p_1 Su_0 - z, \zeta \rangle \geqq 0 \qquad \forall z \in K. \tag{2.2}$$

This lemma will be needed for the error estimates. By calculating the derivative f', it is possible to characterize u_0 more explicitly. We have

$$\langle u, f'(u_0) \rangle = 2\langle u, (p_2 S)^*(p_2 Su_0 - y_T) \rangle + 2\nu \langle u, u_0 \rangle_W \quad \forall u \in U. \tag{2.3}$$

This implies

$$f''(u_0)(u, u) \geqq 2\nu \langle u, u \rangle_W \quad \forall u \in U. \tag{2.4}$$

Hence Lemma 2.2 can also be formulated in the following way.

LEMMA 2.3. *Suppose that* (SL) *is fulfilled. Then $u_0 \in U$ is optimal for* (P) *if and only if u_0 is feasible and there is a $\zeta \in Z^*$ such that for $w := S^*(\zeta, 2(p_2 Su_0 - y_T))$ the inequalities* (2.2) *and*

$$\langle u - u_0, w + 2\nu u_0 \rangle \geqq 0 \quad \forall u \in U_{ad} \tag{2.5}$$

are fulfilled.

We now introduce abstract discretizations for (P). To this end, let for every $i \in \mathbb{N}$ closed convex sets $U_{ad}^i \subset U$, $K_i \subset Z$ and an operator $S_i \in \mathcal{L}(U, Z \times E)$ with the property $S_i^*(Z^* \times E^*) \subset \tilde{U}$ be given. Then we define

($\hat{\mathrm{P}}_i$) Minimize $\| p_2 S_i u - y_T \|_E^2 + \nu \| u \|_W^2$

subject to

$u \in U_{ad}^i, \qquad p_1 S_i u_i \in K_i.$

For every $i \in \mathbb{N}$ let

$$D_i := \{ u \in U \mid p_1 S_i u \in K_i \},$$

and

$$D := \{u \in U \mid p_1 S_u \in K\}.$$

The proof of the following lemma is based on the proof of Theorem 2 in Robinson [15].

LEMMA 2.4. *Suppose that the following conditions are fulfilled.*

(i) *For the element* $\bar{u}$ *of* (SL) *there exists a sequence* $\{\bar{u}_i\}_{i\in\mathbb{N}}$ *such that* $\bar{u}_i \in U_{ad}^i$ *for all* $i \in \mathbb{N}$, $\lim_{i\to\infty} \|\bar{u}_i - \bar{u}\|_U = 0$ *and* $\lim_{i\to\infty} \|p_1 S_i \bar{u}_i - p_1 S\bar{u}\|_Z = 0$.

(ii) $K \subset K_i$ *for all* $i \in \mathbb{N}$.

Let now $u_0 \in U_{ad} \cap D$ *and a sequence* $\{u_i\}_{i\in\mathbb{N}}$ *with* $\lim_{i\to\infty} \|u_i - \bar{u}\|_U = 0$ *and* $u_i \in U_{ad}$ *for all* $i \in \mathbb{N}$ *be given. Then there are* $i_0 \in \mathbb{N}$, $c > 0$ *such that for all* $i \geqq i_0$ *there exists a* $v_i \in U_{ad}^i \cap D_i$ *with*

$$\|v_i - u_0\|_U \leqq \|u_i - u_0\|_U + c\|p_1 S_i u_i - p_1 S u_0\|_Z.$$

Proof. Let $\hat{S} := p_1 S$, $\hat{S}_i := p_1 \hat{S}_i$. By assumption (i) there are $\mu > 0$, $i_0 \in \mathbb{N}$ with $\hat{S}\bar{u} - \mu B \subset K$ and

$$\|\hat{S}_i \bar{u}_i - \hat{S}\bar{u}\|_Z \leqq \frac{\mu}{2} \quad \forall i \geqq i_0$$

(B denotes the unit ball of Z). This implies

$$\hat{S}_i \bar{u}_i - \frac{\mu}{2} B \subset S\bar{u} - \mu B \subset K \quad \forall i \geqq i_0.$$

Thus, from (ii) we get for $\eta := \dfrac{\mu}{2}$

$$\hat{S}_i \bar{u}_i - \eta B \subset K_i \quad \forall i \geqq i_0.$$

Now let $u_0 \in U_{ad} \cap D$ and a sequence $\{u_i\}_{i\in\mathbb{N}}$ such that $u_i \in U_{ad}^i$ for all $i \in \mathbb{N}$ and $\lim_{i\to\infty} \|u_i - u_0\|_U = 0$ be given. Let $i \geqq i_0$. If $\hat{S}_i u_i \in K_i$ define $v_i := u_i$. If $\hat{S}_i u_i \notin K_i$, define $d_i := d(\hat{S}_i u_i, K_i)$. Then for arbitrary $\delta > 0$ there is a $k_\delta \in K_i$ such that for $z_\delta := \hat{S}_i u_i - k_\delta$ the inequalities

$$0 < \|z_\delta\|_Z < d_i + \delta$$

hold. For $\varepsilon \in]0, \delta[\cap]0, \eta[$ we define $z_\varepsilon := -(\eta - \varepsilon)\|z_\delta\|_Z^{-1} z_d$. It follows $\|z_\varepsilon\|_Z = \eta - \varepsilon < \eta$, thus $z_\varepsilon \in \eta B$. Hence there is a $k_\varepsilon \in K_i$ such that $z_\varepsilon = \hat{S}_i \bar{u}_i - k_\varepsilon$. With $\lambda := [1 + (\eta - \varepsilon)\|z_\delta\|_Z^{-1}]^{-1}$ we obtain $0 < \lambda < 1$ and

$$\begin{aligned}\hat{S}_i((1-\lambda)u_i - \lambda\bar{u}_i) - ((1-\lambda)k_\delta - \lambda k_\varepsilon) &= (1-\lambda)(\hat{S}_i u_i - k_\delta) + \lambda(\hat{S}_i \bar{u}_i - k_\varepsilon) \\ &= (1-\lambda)z_\delta + \lambda z_\varepsilon = 0.\end{aligned}$$

For $v_i = (1-\lambda)u_i + \lambda\bar{u}_i$, $k_i = (1-\lambda)k_\delta + \lambda k_\varepsilon$, this implies $v_i \in U_{ad}^i$, $\hat{S}_i v_i = k_i \in K_i$ and therefore $v_i \in U_{ad}^i \cap D_i$. Further, we have

$$\|u_i - v_i\|_U = \|u_i - (1-\lambda)u_i - \lambda\bar{u}_i\|_U = \lambda\|u_i - \bar{u}_i\|_U.$$

Since the sequences $\{u_i\}_{i\in\mathbb{N}}$, $\{\bar{u}_i\}_{i\in\mathbb{N}}$, are bounded, there is a $c > 0$ with

$$\|u_i - \bar{u}_i\|_U \leqq c \quad \forall i \geqq i_0.$$

Because of $\lambda \leqq (\eta - \varepsilon)^{-1}\|z_\delta\|$, $\|z_\delta\| \leqq d_i + \delta$ we get

$$\|v_i - u_0\|_U \leqq \|u_i - u_0\|_U + \lambda\|u_i - \bar{u}_i\|_U \leqq \|u_i - u_0\|_U + \frac{c}{\eta - \varepsilon}(d(\hat{S}_i u_i, K_i) + \delta).$$

Since $\hat{S}u_0 \in K \subset K_i$ the proof is completed by letting δ and ε approach zero. □

LEMMA 2.5. *Suppose that* (SL) *holds. Then there exists* $c>0$ *such that for all* $u\in U_{ad}$ *there is a* $\tilde{u}\in U_{ad}\cap D$ *with*

$$\|u-\tilde{u}\|_U \leqq c\|u-\bar{u}\|_U\, d(p_1Su, K).$$

Proof. Define the multivalued function $F: U\to Z$ by

$$F(u) := \begin{cases} p_1 Su - K & \text{if } u\in U_{ad},\\ \varnothing & \text{if } u\notin U_{ad}.\end{cases}$$

Then F is a closed convex function, and by the assumptions of the lemma there exists $\eta>0$ with $\eta B\subset F(\bar{u})$. The assertion of the lemma is therefore a special case of Theorem 2 in Robinson [15]. □

The following Proposition is an immediate consequence of Lemma 2.1 and Lemma 2.4.

PROPOSITION 2.1. *Let the assumptions of Lemma* 2.1 *and Lemma* 2.4 *be fulfilled. Then there is an* $i_0\in\mathbb{N}$ *such that for all* $i\geqq i_0$ *the discrete problem* $(\hat{P}_i)$ *has an optimal solution.*

In the following we denote by c a generic constant. Let

$$f_i(u) := \|p_2S_iu - y_T\|_E^2 + \nu\|u\|_W^2 \quad \forall u\in U.$$

THEOREM 2.1. *In addition to the assumptions of Lemma* 2.1 *and Lemma* 2.4, *suppose that the following assertions hold.*

(i) $U_{ad}^i\subset U_{ad}\quad \forall i\in\mathbb{N}$.

(ii) *If* $M\subset U_{ad}$ *is bounded then* $\{S_iu \mid u\in M\}$ *is also bounded.*

Now let $\{w_i\}_{i\in\mathbb{N}}$ *be a sequence with* $w_i\in U_{ad}^i$ *for all* $i\in\mathbb{N}$ *and* $\lim_{i\to\infty}\|w_i-u_0\|_U=0$. *Then there are* $i_0\in\mathbb{N}$, $c>0$, *and a bounded sequence* $\{v_i\}_{i\geqq 0}$ *with* $v_i\in U_{ad}^i$, $p_1S_iv_i\in K_i$ *for all* $i\geqq i_0$ *such that the following error estimates hold for the optimal solutions* u_0 *of* $(\hat{P})$ *and* u_i *of* $(\hat{P}_i)$ $(i\geqq i_0)$:

(a) $f_i(u_i)-f(u_0)\leqq c\|p_2S_iv_i-p_2Sv_i\|_E + c\|p_1S_iw_i-p_1Sw_i\|_Z + c\|w_i-u_0\|_U$,

(b) $f(u_0)-f_i(u_i)\leqq c\|S_iu_i-Su_i\|_{Z\times E} + c\,d(p_1S_iu_i, K)$,

(c) $\nu\|u_i-u_0\|_W^2\leqq f_i(u_i)-f(u_0)+c\|S_iu_i-Su_i\|_{Z\times E}+c\,d(p_1S_iu_i, K)$.

Proof. Suppose $u_1, u_2\in U$ and define

$$\sigma := \max\{\|u_1\|_W, \|u_2\|_W, \|p_2S_iu_1\|_E, \|p_2S_iu_2\|_E\}.$$

Then it can be easily seen that

$$|f_i(u_1)-f(u_2)|\leqq c\sigma\|p_2S_iu_1-p_2Su_2\|_E + c\sigma\|u_1-u_2\|_W. \tag{2.6}$$

From Lemma 2.4 we deduce the existence of an $i_0\in\mathbb{N}$ such that for all $i\geqq i_0$ there is a $v_i\in U_{ad}^i$ with $p_1S_iv_i\in K_i$ and

$$\|v_i-u_0\|_U\leqq\|w_i-u_0\|_U + c\|p_1S_iw_i-p_1Su_0\|_Z.$$

This implies in particular that the sequence $\{v_i\}_{i\geqq i_0}$ is bounded. Thus, from (2.6) we get

$$|f_i(v_i)-f(u_0)|\leqq c\|p_2S_iv_i-p_2Su_0\|_E + c\|v_i-u_0\|_U.$$

Hence we obtain

$$\begin{aligned} f_i(u_i)-f(u_0) &\leqq f_i(v_i)-f(u_0)\\ &\leqq c\|p_2S_iv_i-p_2Sv_i\|_E + c\|v_i-u_0\|_U\\ &\leqq c\|p_2S_iv_i-p_2Sv_i\|_E + c\|p_1S_iw_i-p_1Su_0\|_Z + c\|w_i-u_0\|_U\\ &\leqq c\|p_2S_iv_i-p_2Sv_i\|_E + c\|p_1S_iw_i-p_1Sw_i\|_Z + c\|w_i-u_0\|_U. \end{aligned}$$

This shows inequality (a).

Lemma 2.5 shows that for every $i \in \mathbb{N}$ there is a $\tilde{u}_i \in U_{ad}$ with $p_1 S\tilde{u}_i \in K$ and

$$\|u_i - \tilde{u}_i\|_U \leqq c\|u_i - \bar{u}\|_U \, d(p_1 S u_i, K).$$

Thus, since U_{ad} is bounded, the sequence $\{u_i\}_{i \geqq i_0}$ is bounded, and we get

$$\|u_i - \tilde{u}_i\|_U \leqq c\, d(p_1 S u_i, K) \leqq c\|p_1 S_i u_i - p_1 S u_i\|_Z + c\, d(p_1 S_i u_i, K). \tag{2.7}$$

In particular $\{\tilde{u}_i\}_{i \geqq i_0}$ is bounded. Hence (2.6) shows

$$|f_i(u_i) - f(\tilde{u}_i)| \leqq c\|p_2 S_i u_i - p_2 S\tilde{u}_i\|_E + c\|u_i - \tilde{u}_i\|_U.$$

Thus we get the estimates

$$\begin{aligned} f(u_0) - f_i(u_i) &\leqq f(\tilde{u}_i) - f_i(u_i) \\ &\leqq c\|p_2 S_i u_i - p_2 S\tilde{u}_i\|_E + c\|u_i - \tilde{u}_i\|_U \\ &\leqq c\|p_2 S_i u_i - p_2 S u_i\|_E + c\|u_i - \tilde{u}_i\|_U \\ &\leqq c\|S_i u_i - S u_i\|_{Z \times E} + c\, d(p_1 S_i u_i, K). \end{aligned}$$

It remains to show (c). Let ζ be as in Lemma 2.2. Then we get from (2.1) and (2.4)

$$\begin{aligned} &f(u_i) + \zeta p_1 S(u_i) - (f(u_0) + \zeta p_1 S(u_0)) \\ &\quad = (f'(u_0) + (p_1 S)^* \zeta)(u_i - u_0) + \tfrac{1}{2} f''(u_0)(u_i - u_0, u_i - u_0) \geqq \nu \|u_i - u_0\|_W^2. \end{aligned}$$

Because of $p_1 S\tilde{u}_i \in K$, the inequalities (2.2), (2.6), (2.7) imply

$$\begin{aligned} &f(u_i) + \zeta p_1 S(u_i) - (f(u_0) + \zeta p_1 S(u_0)) \\ &\quad = f(u_i) - f(u_0) + \zeta p_1 S(u_i) - \zeta p_1 S(\tilde{u}_i) - \zeta(p_1 S u_0 - p_1 S\tilde{u}_i) \\ &\quad \leqq f(u_i) - f(u_0) + \zeta p_1 S(u_i - \tilde{u}_i) \\ &\quad \leqq f(u_i) - f_i(u_i) + f_i(u_i) - f(u_0) + c\|u_i - \tilde{u}_i\|_U \\ &\quad \leqq f_i(u_i) - f(u_0) + c\|p_2 S_i u_i - p_2 S u_i\|_E + c\|p_1 S_i u_i - p_1 S u_i\|_Z + c\, d(p_1 S_i u_i, K). \end{aligned}$$ □

We just mention the special case that ($\hat{\mathrm{P}}$) contains no state constraint.

COROLLARY 2.1. *Suppose that $K = Z$ and let the assumptions of Lemma* 2.1 *and assumptions* (i) *and* (ii) *of Theorem* 2.1 *be fulfilled. Let the sequence $\{w_i\}_{i \in \mathbb{N}}$ be as in Theorem* 2.1. *Then the following estimates hold*:

(a) $f_i(u_i) - f(u_0) \leqq c\|p_2 S_i w_i - p_2 S w_i\|_E + c\|w_i - u_0\|_U$,

(b) $f(u_0) - f_i(u_i) \leqq c\|p_2 S_i u_i - p_2 S u_i\|_E$,

(c) $\nu\|u_i - u_0\|_W^2 \leqq f_i(u_i) - f(u_0) + c\|p_2 S_i u_i - p_2 S u_i\|_E$.

Remark. If (P) has neither a state nor a control constraint, then estimate (a) of Corollary 2.1 can be sharpened. We get

$$f_i(u_i) - f(u_0) \leqq c\|p_2 S_i w_i - p_2 S w_i\|_E + c\|w_i - u_0\|_W^2. \tag{2.8}$$

The proof of (2.8) is an application of Lemma 2.2 and (2.6); we get

$$\begin{aligned} f_i(u_i) - f(u_0) &\leqq f_i(w_i) - f(w_i) + f(w_i) - f(u_0) \\ &\leqq f_i(w_i) - f(w_i) + f'(u_0)(w_i - u_0) + \tfrac{1}{2} f''(u_0)(w_i - u_0, w_i - u_0) \\ &\leqq c\|p_2 S_i w_i - p_2 S w_i\|_E + c\|w_i - u_0\|_W^2. \end{aligned}$$

Together with (c) of Corollary 2.1 this implies

$$\|u_i - u_0\|_W^2 \leqq c\|p_2 S_i u_i - p_2 S u_i\|_E + c\|p_2 S_i w_i - p_2 S w_i\|_E + c\|w_i - u_0\|_W^2. \tag{2.9}$$

The square in the last term of (2.9) can also be obtained if a control constraint is present (cf. Malanowski [14]), but obviously an estimate of this kind is limited to the coercive case without state constraints.

Remark. From Theorem 2.1 we easily obtain an abstract convergence theorem for the discrete optimal values and optimal solutions; but instead of applying such a result in a concrete situation, it is more convenient to use the estimates of Theorem 2.1 directly (compare, e.g., the proof of Theorem 3.1).

3. Error estimates for discretizations of the parabolic optimal control problem. In the first part of this section we define discretizations for the problem (P). We start with the partial differential equation for $\beta \neq 0$. In this case the variational form of (1.1)-(1.3) may be used in the usual way. To this end define

$$a(v, w) := \sum_{i,j=1}^{n} \left(a_{ij} \frac{\partial v}{\partial x_i}, \frac{\partial w}{\partial x_j} \right) + (a_0 v, w) + \alpha \beta^{-1} (v|_\Gamma, w|_\Gamma) \quad \forall u, w \in H^1(\Omega).$$

Equations (1.1), (1.2) are then equivalent to

$$\frac{d}{dt}(v, y(t)) + a(v, y(t)) = (v|_\Gamma, u(t)) \quad \forall v \in H^1(\Omega), \quad \forall' t \in [0, T]. \tag{3.1}$$

Let for all $h \in]0, h_0]$, $(h_0 > 0)$ a finite dimensional subspace V_h of $H^1(\Omega)$, be given. Then there is a unique function y_h such that $y_h(t) \in V_h$ for all $t \in [0, T]$, $y_h(0) = 0$, and

$$\frac{d}{dt}(v, y_h(t)) + a(v, y_h(t)) = (v|_\Gamma, u(t)) \quad \forall v \in V_h, \quad \forall' t \in [0, T]. \tag{3.2}$$

For the subspaces V_h we make the following assumptions ($\|\cdot\|_s$ denotes the norm of $H^s(\Omega)$, respectively, $H^s(\Gamma)$):

$$\inf_{w \in V_h} \{\|v - w\| + h\|v - w\|_1\} \leqq c_s h^s \|v\|_s \quad \forall v \in H^s(\Omega), \quad \forall s \in [1, 2]; \tag{3.3}$$

$$\|w\|_1 \leqq c h^{-1} \|w\| \quad \forall w \in V_h; \tag{3.4}$$

$$\text{the family } \{V_h\}_{h \in]0, h_0]} \text{ is dense in } L^2(\Omega). \tag{3.5}$$

An example for such spaces is given by the usual piecewise linear finite element spaces.

For problem (P) the operator S of § 2 is given by

$$Su := (y(u), y(u)(T)) \quad \forall u \in U,$$

where $y(u)$ solves (1.1)-(1.3). For the space U we may choose

$$U := L^p(0, T; L^2(\Gamma)) \tag{3.6}$$

with $p = 2$ if $\beta \neq 0$ and $p > 4$ if $\beta = 0$ (cf. Washburn [16], e.g.). As we have seen above, for $\beta \neq 0$ an approximation S_h^1 of S can be defined by

$$p_1 S_h^1(u) := y_h(u) \quad \text{for all } u \in U.$$

Now let $\beta = 0$. Then an application of the finite element method is much more complicated since the variational form cannot be used and since the controls are irregular. It is possible to overcome these difficulties by applying the input formula

$$y(u)(t) = \int_0^t AS(t - \tau) Du(\tau)\, d\tau \tag{3.7}$$

as it is shown in Lasiecka [8]. We can only give a short sketch of this method and refer the reader to Lasiecka [8] for all further details.

In (3.7) $S(t)$ denotes the semigroup generated by A and D denotes the Dirichlet map

$$Dv := w, \quad \text{where } Aw = 0,\ w|_\Gamma = v.$$

The operators A, $S(t)$, D are now approximated separately. To this end, let $\{V_h\}_{h\in]0,h_0]}$ be a family of subspaces of $H^1(\Omega)$ with the following properties:

(3.8) $\{V_h\}$ is an $S_h^{4,2}(\Omega)$-system;

(3.9) $$V_h|_\Gamma \subset H^1(\Gamma);$$

(3.10) $$\inf_{w\in V_h} \{\|v-w\| + h\|v-w\|_1 + h^{1/2}\|v|_\Gamma - w|_\Gamma\| + h^{3/2}\|v|_\Gamma - w|_\Gamma\|_1\} \leqq ch^s\|v_s\| \quad \forall v\in H^s(\Omega),\ \forall s\in[2,4];$$

(3.11) $$\left\|\frac{\partial v}{\partial n_A}\right\| \leqq ch^{-1/2}\|v\|_1 \quad \forall v\in V_h;$$

(3.12) $$\|v|_\Gamma\| \leqq ch^{1/2}\|v\|_1 \quad \forall v\in V_h;$$

(3.13) $$\|v\|_1 \leqq ch^{-1}\|v\| \quad \forall v\in V_h.$$

Define (with suitable $\gamma>0$) $A_h\colon V_h\to V_h$ by

$$(A_h v, w) := a(v,w) - \left(v|_\Gamma, \frac{\partial w}{\partial n_A}\right) - \left(\frac{\partial v}{\partial n_A}, w|_\Gamma\right) + \gamma h^{-1}(v|_\Gamma, w|_\Gamma) \quad \forall v, w\in V_h.$$

Let $\{\tilde U_h^\Gamma\}_{h\in]0,h_0]}$ be a family of spaces of piecewise constant functions on Γ and define $D_h\colon \tilde U_h^\Gamma \to V_h$ by

$$(-AD_h v, Aw) + h^{-3}(v|_\Gamma - (D_h v)|_\Gamma, w|_\Gamma) = 0 \quad \forall w\in V_h.$$

Then, for $u_h\in L^p(0,T;\tilde U_h^\Gamma)$ with

$$S_h(t) := e^{-A_h t} \quad \forall t>0$$

an approximation of $y(u_h)$ can be introduced by

$$y_h(u_h)(t) = \int_0^t A_h S_h(t-\tau) D_h u_h(\tau)\, d\tau.$$

Denote by $\tilde P_h$ the orthogonal projection of $L^2(\Gamma)$ onto $\tilde U_h^\Gamma$. Then, for $\beta = 0$ we define an approximation S_h^2 of S by

$$p_1 S_h^2 u = y_h(\tilde P_h u) \quad \forall u\in U.$$

In the case of $n=1$ and $\alpha\neq 0$, $\beta\neq 0$, the Fourier series of $y(u)$, can be used to derive a practically useful approximation of S. We have for all $u\in L^q(0,T)$ with $q>2$

(3.14) $$y(u)(t) = \sum_{k=1}^{\infty} v_k(1) \int_0^t e^{-\lambda_k(t-\tau)} u(\tau)\, d\tau\, v_k.$$

Here, λ_k (respectively, v_k) denote the eigenvalues (respectively, eigenfunctions) of the corresponding elliptic eigenvalue problem. We define $y_m(u)$ by replacing "∞" in (3.14) by "m" and

$$p_1 S_m^3 u = y_m(u) \quad \forall u\in L^q(0,T).$$

We remark that $y_m(u)$ may be viewed as a discretization in t and x since, at least for simple functions u, the integrals in (3.14) can be evaluated explicitly.

It is a well-known fact that in the above situation the functions v_k are uniformly bounded with respect to the maximum norm and that the eigenvalues λ_k behave asymptotically as k^2. Thus we have for $u \in L^\infty(0, T)$

$$\|y_m(u)(t) - y(u)(t)\|^2 \leqq \sum_{k=m+1}^{\infty} \left| \int_0^t e^{-\lambda_k(t-\tau)} u(\tau)\, d\tau \right|^2 \leqq \|u\|_\infty^2 \sum_{k=m+1}^{\infty} \left| \int_0^t e^{-\lambda_k(t-\tau)}\, d\tau \right|^2$$

$$\leqq c\|u\|_\infty^2 \sum_{k=m+1}^{\infty} \frac{1}{\lambda_k^2} \leqq c\|u\|_\infty^2 \int_{m+1}^{\infty} \frac{1}{x^4}\, dx \leqq c\|u\|_\infty^2 m^{-3},$$

and finally

$$\|y_m(u)(t) - y(u)(t)\| \leqq c m^{-3/2} \|u\|_\infty. \tag{3.15}$$

In the following Proposition, ε denotes an arbitrary small positive number.

PROPOSITION 3.1. (a) *If $\beta \neq 0$ and assumptions* (3.3)-(3.5) *are satisfied then for all $u \in L^\infty(0, T; L^2(\Gamma))$, the following estimate holds*:

$$\|p_1 Su(t) - p_1 S_h^1 u(t)\| \leqq c h^{3/2-\varepsilon} \|u\|_{L^\infty(0,T;L^2(\Gamma))} \quad \forall t \in [0, T].$$

(b) *If $\beta = 0$ and assumptions* (3.8)-(3.13) *are satisfied, then for all $u \in L^\infty(0, T; \tilde{U}_h^\Gamma)$ the following estimate holds*:

$$\|p_1 Su(t) - p_1 S_h^2 u(t)\| \leqq c h^{1/2-\varepsilon} \|u\|_{L^\infty(0,T;L^2(\Gamma))} \quad \forall t \in [0, T].$$

(c) *In the case of $\alpha \neq 0$, $\beta \neq 0$, $n = 1$ we have for every $u \in L^\infty(0, T)$*

$$\|p_1 Su(t) - p_1 S_m^3 u(t)\| \leqq c m^{-3/2} \|u\|_{L^\infty(0,T)} \quad \forall t \in [0, T].$$

Proof. Assumption (a) is only a slight modification of Proposition 1 in Knowles [5].
Assumption (b) follows from Theorem 2.2.1 of Lasiecka [8].
Assumption (c) was shown in (3.15). □

We can now apply Theorem 2.1. Let U be as given in (3.6). The choice of the sets Z, E, W, $\tilde{U}$, U_{ad} is obvious. Let K be defined by $K := \{z \in Z \mid z(t) \in C(t)\ \forall t \in [0, T]\}$. Let $\{U_h^\Gamma\}_{h \in]0, h_0]}$ be a family of subspaces of $L^2(\Gamma)$ with the following properties:

(3.16) For every $h \in]0, h_0]$ the set U_h^Γ is a space of piecewise constant functions on Γ. The family $\{U_h^\Gamma\}_h$ is dense in $L^2(\Gamma)$ and $h_1 > h_2$ implies $U_{h_1}^\Gamma \subset U_{h_2}^\Gamma$.

Let p be as in the definition of U and define $U_{ad}^h := U_h \cap U_{ad}$ with $U_h := L^p(0, T; U_h^\Gamma)$. In the case of $b = 0$, we choose $\tilde{U}_h^\Gamma = U_h^\Gamma$. Further, let an approximation $C_h(t)$ of $C(t)$ be given such that

$$C(t) \subset C_h(t) \quad \forall h \in]0, h_0], \quad \forall t \in [0, T]. \tag{3.17}$$

K_h is defined analogously to K. Denote by P_t the projection of $L^2(\Omega)$ onto $C(t)$, and suppose that for $j = 1, 2$ and $X := \{p_1 S_h^j u(t) \mid u \in U_{ad},\ t \in [0, T],\ h \in]0, h_0]\}$

$$\lim_{h \to 0} \sup_{t \in [0,T]} \sup_{v \in X \cap C_h(t)} \|v - P_t v\| = 0. \tag{3.18}$$

In this way we have obtained discretizations (P_h) of (P). Herein, S is approximated as in (a), respectively, (b) of Proposition 2.1. We still mention that (P) has an optimal solution u_0.

THEOREM 3.1. *Suppose that for* (P) *the Slater condition* (SL) *is fulfilled and that* (3.3)-(3.5), *respectively,* (3.8)-(3.13) *holds. In addition let* (3.16)-(3.18) *be satisfied. Then the following assertions are valid.*

(a) *There is* $h_1 \in]0, h_0]$ *such that problem* (P_h) *has an optimal solution* u_h *for every* $h \in]0, h_1]$.

(b) $\lim_{h\to 0} \min (P_h) = \min (P)$.

(c) $\lim_{h\to 0} \|u_h - u_0\| = 0$ *if* $\nu > 0$.

Proof. By (3.16) there is a sequence $\{\hat{w}_h\}_{h>0}$ in U with $\hat{w}_h \in U_h$ for every $h > 0$ and $\lim_{h\to 0} \|\hat{w}_h - u_0\| = 0$. Let Q be the projection of $L^2(\Gamma)$ onto U_{ad} and define $w_h := Q\hat{w}_h$. Since Q is Lipschitz continuous and $u_0 \in U_{ad}$, we have $w_h \in U_{ad}$ and $\lim_{h\to 0} \|w_h - u_0\| = 0$. In the same way, it follows that there is a sequence $\{\bar{u}_h\}_{h>0}$ with $\bar{u}_h \in U_{ad}^h$ for every $h > 0$ and $\lim_{h\to 0} \|\bar{u}_h - \bar{u}\| = 0$. Thus, by Proposition 3.1 and the fact that $\|\bar{u}_h\|_{L^\infty(\Sigma)} \leqq \max\{\rho_1, \rho_2\}$ for all $h > 0$, we see that assumption (i) of Lemma 2.4 is fulfilled. Hence (3.17) and Proposition 2.1 imply (a).

Assumption (i) of Theorem 2.1 is satisfied by the definition of U_{ad}^h, and assumption (ii) of Theorem 2.1 follows from Proposition 3.1. Thus we can apply the estimates (a)-(c) of Theorem 2.1. The terms containing S and S_h^j $(j = 1, 2)$ converge to zero again by Proposition 3.1 and the uniform boundedness of the sequences $\{v_h\}$, $\{w_h\}$, $\{u_h\}$ with respect to the norm of $L^\infty(\Sigma)$. Hence assertions (b) and (c) are shown if $\lim_{h\to 0} d(p_1 S_h^j u_h, K) = 0$, but this is an immediate consequence of (3.18). □

We briefly discuss assumption (3.18) for a typical example. Let

$$C := \{v \in L^2(\Omega) \,|\, (v, g_\nu) \leqq \alpha_\nu,\ \nu = 1, \cdots, l\} \tag{3.19}$$

with $g_\nu \in L^2(\Omega)$, $\alpha_\nu \in \mathbb{R}$, $\nu = 1, \cdots, l$. Under the assumptions of Proposition 3.1, the set X is bounded by $r > 0$. With $g_{\nu h} \in L^2(\Omega)$ and

$$\varepsilon_{\nu h} := r\|g_{\nu h} - g_\nu\| \tag{3.20}$$

we define

$$C_h := \{v \in L^2(\Omega) \,|\, (v, g_{\nu h}) \leqq \alpha_\nu + \varepsilon_{\nu h},\ \nu = 1, \cdots, l\}. \tag{3.21}$$

LEMMA 3.1. *Suppose that* $\|g_{\nu h} - g_\nu\| \leqq ch^\gamma$, $\nu = 1, \cdots, l$, *with* $\gamma > 0$. *Let the sets* C, C_h *be defined by* (3.19), (3.20), *and assume there exists* $w \in C$ *with* $(w, g_\nu) < \alpha_\nu$ *for all* $\nu \in \{1, \cdots, l\}$. *Then* $X \cap C \subset X \cap C_h$ *and*

$$\sup_{v \in X \cap C_h} \|v - Pv\| \leqq ch^\gamma,$$

where P *denotes the projection of* $L^2(\Omega)$ *onto* C.

Proof. The relation $X \cap C \subset X \cap C_h$ is an immediate consequence of the definition of $\varepsilon_{\nu h}$ and the fact that

$$(v, g_{\nu h}) \leqq \|v\|\,\|g_{\nu h} - g_\nu\| + (v, g_\nu).$$

Now let $v \in X \cap C_h$ be given with $v \notin C$. Then $I = \{\nu \,|\, (v, g_\nu) > \alpha_\nu\} \neq \varnothing$. For $\nu \in I$ define

$$\lambda_\nu = ((v, g_\nu) - \alpha_\nu)((v, g_\nu) - (w, g_\nu))^{-1},$$

and

$$\lambda = \max_{\nu \in I} \lambda_\nu, \quad \mu = \min_{\nu \in I} \{\alpha_\nu - (w, g_\nu)\}, \quad z = v + \lambda(w - v).$$

This implies $0 < \lambda < 1$, $\mu > 0$,

$$(z, g_\nu) = (v, g_\nu) + \lambda((w, g_\nu) - (v, g_\nu)) \leqq (v, g_\nu) + \lambda_\nu((w, g_\nu) - (v, g_\nu)) = \alpha_\nu$$

for all $\nu \in I$, and

$$(z, g_\nu) = (v, g_\nu) + \lambda((w, g_\nu) - (v, g_\nu)) \leqq (v, g_\nu) \leqq \alpha_\nu$$

for all $\nu \notin I$. Hence $z \in C$ and therefore

$$\|v - Pv\| \leqq \|v - z\| = \lambda \|w - v\| \leqq 2r\lambda.$$

Let $\nu \in I$ such that $\lambda = \lambda_\nu$. Then

$$\begin{aligned}\lambda &\leqq ((v, g_\nu) - \alpha_\nu)\mu^{-1} = ((v, g_{\nu h}) - \alpha_\nu + (v, g_\nu) - (v, g_{\nu h})\mu^{-1} \\ &\leqq (\varepsilon_{\nu h} + r\|g_\nu - g_{\nu h}\|)\mu^{-1} \leqq 2\mu^{-1} r h^\gamma,\end{aligned}$$

which completes the proof. □

Remark. In Lemma 3.1 we have only shown $X \cap C \subset X \cap C_h$. Since we can replace C by $X \cap C$, respectively, C_h by $X \cap C_h$ without changing (P) respectively, (P_h), we get $C \subset C_h$. Furthermore, it is obvious that this lemma can be easily extended to the more general situation where the functions g_ν and α_ν are time dependent.

We conclude this section by paying some attention to the case $n = 1$ where we obtain convergence rates due to the fact that only semidiscretizations are considered.

Let $\alpha \neq 0$, $\beta \neq 0$, let S be approximated by S_m^3, and let C, C_h be defined by (3.19)–(3.21) (where we write $g_{\nu m}$, $\varepsilon_{\nu m}$ instead of $g_{\nu h}$, $\varepsilon_{\nu h}$). This leads to a semidiscretized version (P_m) of (P), since the set of feasible controls requires no approximation here. The next theorem follows directly from Theorem 2.1 and Lemma 3.1.

THEOREM 3.2. *Let* (P), (P_m) *be as described above. Suppose that in addition to the Slater condition* (SL) *the following condition holds*:

$$\|g_{\nu m} - g_\nu\| \leqq cm^{-1}, \quad \nu = 1, \cdots, l.$$

Then for sufficiently large m the problem (P_m) *has an optimal solution* u_m. *Furthermore, the following estimate holds*:

(a) $\|\min(P_m) - \min(P)\| \leqq cm^{-1}$,

(b) $\|u_m - u_0\| \leqq cm^{-1/2}$ *if* $\nu > 0$.

Remark. 1. It is clear that Theorem 3.2 also holds for the situations described in Proposition 3.1(a), (b) (and $n = 1$).

2. As we have already mentioned, S_m^3 can be viewed as a discretization of S in t and x. This fact can be easily used to formulate a convergence theorem for a fully discretized problem in the case of one-space dimension.

4. The bang–bang case. The results of the preceding section contain no information about the convergence of the discrete optimal controls if $\nu = 0$. As we shall see, results in this direction require some knowledge about the qualitative behaviour of the optimal controls of the continuous problem (P). This is obtained by analyzing the optimality conditions. For simplicity we assume in this section that $\rho_2 = -\rho_1 = \rho > 0$.

Let $NBV(0, T; L^2(\Omega))$ be the space of all functions $v: [0, T] \to L^2(\Omega)$ vanishing at T which are of bounded variation and right continuous. It is well known that each functional $\zeta \in C([0, T]; L^2(\Omega))^*$ can be uniquely represented by a function $v_\zeta \in NBV(0, T; L^2(\Omega))$. Using this fact, the following lemma is a direct consequence of Lemma 2.3.

LEMMA 4.1. *Suppose* $\nu = 0$ *and let* (SL) *hold. Then* $u_0 \in U$ *is optimal for* (P) *if and only if* u_0 *is feasible and there is a functional* $\zeta \in C([0, T]; L^2(\Omega))^*$ *such that for* $w := S^*(\zeta, 2(p_2 Su_0 - y_T))$ *the following equations are fulfilled*:

$$u_0 w + \rho|w| = 0, \tag{4.1}$$

$$\sup_{z \in K} \int_0^T z(t)\, dv_\zeta = \int_0^T p_1 Su_0(t)\, dv_\zeta. \tag{4.2}$$

In order to get results on the structure of u_0, the equations of Lemma 4.1 must be analyzed in detail. This requires at first a precise description of S^*. To this end we quote some results of Mackenroth [9].

Each $z^* \in Z^*(Z = C([0, T]; L^2(\Omega)))$ can be written in the form

$$\langle z, z^* \rangle = (z(0), \zeta_0) + (z(T), \zeta_T) + \langle z, \hat{\zeta} \rangle \qquad \forall z \in Z \tag{4.3}$$

with $\zeta_0, \zeta_T \in L^2(\Omega)$ and a functional $\hat{\zeta} \in Z^*$ such that $v_{\hat{\zeta}}$ is continuous at zero and T (cf. Mackenroth [9, Satz 5.1]). Moreover, $\hat{\zeta}$ can be viewed as an element of $\mathcal{D}^*(]0, T[; V^*)$ (the space of distributions on $]0, T[$ with values in V^*) where

$$V := \left\{ v \in H^1(\Omega) \,|\, \alpha v|_\Gamma + \beta \frac{\partial v}{\partial n_A} = 0 \right\}.$$

Let the bilinear form a be defined as in § 3, but without the boundary term, and define the operator $\mathcal{A}$ by

$$\langle w, \mathcal{A}v \rangle = a(v, w) \quad \forall v, w \in V.$$

Then the following equation is well defined (in the sense of distributions on $\mathcal{D}^*(]0, T[; V^*)$):

$$-\frac{dp}{dt} + \mathcal{A}_p = \hat{\zeta}. \tag{4.4}$$

Moreover, it can be shown that (4.4) and

$$p(T) = 2(p_2 S u_0 - y_T) + \zeta_T \tag{4.5}$$

has a unique solution p with

$$p \in L^2(0, T; V) \cap NBV(0, T; V^*) \cap L^\infty(0, T; L^2(\Omega)) \tag{4.6}$$

(cf. Mackenroth [9, Satz 5.2, Satz 4.2], and Mackenroth [12, Thm. 5]).

LEMMA 4.2. *Let u_0, ζ, w, p be given as in Lemma* 4.1 (*respectively, as in* (4.3)-(4.5)). *Then we have*

$$w = p|_\Sigma, \qquad \text{if } \beta = 1$$

$$w = -\frac{\partial p}{\partial n_A}, \quad \text{if } \alpha = 1, \quad \beta = 0.$$

Proof. Set $S_1 := p_1 S$ and $U := L^\infty(0, T; L^2(\Gamma))$. We shall use the input formula

$$S_1 u(t) = \int_0^t AS(t - \tau) G u(\tau)\, d\tau \quad \forall t \in [0, T], \quad \forall u \in U,$$

where S denotes the semigroup associated with A. The map G is defined by $Gu := w$, where $u \in L^2(\Gamma)$ and $Aw = 0$, $\alpha w|_\Gamma + \beta(\partial w / \partial n_A) = u$. Let $v := v_{\hat{\zeta}}$. Then we get

$$\langle S_1 u, \zeta \rangle = \int_0^T \int_0^t AS(t - \tau) G u(\tau)\, d\tau\, dv = \int_0^T \int_\tau^T AS(t - \tau) G u(\tau)\, dv\, d\tau$$

$$= \int_0^T (u(\tau), \int_\tau^T (AS(t - \tau) G)^*\, dv)\, d\tau.$$

Here we have applied a vector-valued version of the Fubinis Theorem (cf. Dinculeanu [3]). This is possible since we have with $\theta = \frac{1}{4} - \varepsilon$ for $\beta \neq 0$ and $\theta = \frac{3}{4} - \varepsilon$ for $\beta = 0$ (and arbitrary small $\varepsilon > 0$)

$$\int_0^T \int_0^t \|AS(t-\tau)Gu(\tau)\| \, d\tau \, d\|v(t)\| \leqq \|u\|_U \int_0^T \int_0^t (t-\tau)^{-\theta} \, d\tau \, d\|v(t)\|$$

$$\leqq \|u\|_U \int_0^T t^{1-\theta} \, d\|v(t)\| < +\infty$$

(cf. Washburn [16]). Thus

$$S_1^* \hat{\zeta}(t) = \int_t^T (AS(\tau - t)G)^* \, dv. \tag{4.7}$$

Now let $\tilde{S}$ be defined by $\tilde{S}w := (y, y(T))$ where $w \in L^2(0, T; L^2(\Omega))$ with y satisfying (1.3) and

$$\frac{\partial y}{\partial t} + A_y = w, \qquad \alpha y|_\Sigma + \beta \frac{\partial y}{\partial n_A} = 0.$$

Then $\tilde{S}_1 := p_1 \tilde{S}$ can be written as

$$\tilde{S}_1 w(t) = \int_0^t S(t-\tau) w(\tau) \, d\tau.$$

From Mackenroth [9, Satz 5.3], we obtain $\tilde{S}^*(\hat{\zeta}) = q_1$, where q_1 is the solution of (4.4) and $q_1(T) = 0$. A computation quite similar to that made above shows

$$\tilde{S}_1^* \hat{\zeta}(t) = \int_t^T S^*(\tau - t) \, dv. \tag{4.8}$$

Let $Rw := w|_\Gamma$ for $\beta = 1$ and $Rw := -(\partial w / \partial n_A)$ for $\alpha = 1$, $\beta = 0$. Then we have (cf. Washburn [16, p. 664]) $(AS(\tau - t)G)^* = RS^*(\tau - t)$. Thus, using (4.7) and (4.8) we see

$$S_1^* \hat{\zeta} = R \int_t^T S^*(\tau - t) \, dv = Rq_1.$$

Let q_2 be the solution of

$$-\frac{dq_2}{dt} + \mathcal{A} q_2 = 0, \qquad q_2(T) = 2(p_2 S u_0 - y_T) + \zeta_T.$$

It is well known that $Rq_2 = (p_2 S)^*(2(p_2 S u_0 - y_T) + \zeta_T)$. Thus, since $p = q_1 + q_2$ the lemma is shown. □

To make the arguments in the following lemma somewhat simpler, we suppose from now on that $C(t)$ is time independent, i.e., $C(t) = C$, but this is in no way essential. For $u \in U_{ad}$ define

$$M_u := \{t \in [0, T] \mid y(u)(t) \in \operatorname{int} C\}.$$

LEMMA 4.3. *Let the situation be as in Lemma* 4.1. *Then* $v_\zeta|_{M_{u_0}} = 0$.

Proof. Set $y := y(u_0)$, $v := v_\zeta$, and let $]a, b[\subset M_{u_0}$ be given such that $[a, b] \subset M_{u_0}$. Since M_{u_0} is open, there is $\delta_0 > 0$ such that $I_{\delta_0} := [a - \delta_0, b + \delta_0] \subset M_{u_0}$. Let $z \in \dot{Z}$ be arbitrary and choose $w_\delta \in Z$ such that (with $\delta \in]0, \delta_0]$ and $I_\delta := [a - \delta, b + \delta]$)

$$w_\delta(t) := \begin{cases} w(t) & \forall t \in [a, b], \\ 0 & \forall t \in [0, T] \setminus I_\delta, \end{cases}$$

and $\|w_\delta\|_Z \leqq \|w\|_Z$. Denote by B the unit ball of $L^2(\Omega)$. There is a function $\varepsilon(t)$ on I_δ with the property

$$y(t)+\varepsilon(t)B \subset C \quad \forall t \in I_\delta$$

and ε can be chosen such that

$$\varepsilon_0 := \inf_{t\in I_\delta} \varepsilon(t) > 0.$$

For $z := y+\varepsilon_0\|w\|_Z^{-1}w_\delta$, we have $\|z(t)-y(t)\| \leqq \varepsilon_0$ if $t \in I_\delta$ and $z(t)=y(t)\in C$ if $t \in [0, T]\backslash I_\delta$, hence $z(t)\in C$ for all $t\in[0, T]$. Thus (4.2) implies

$$\int_a^b z\,dv \leqq \int_a^b y\,dv + \int_{J_\delta} (y-z)\,dv \leqq \int_a^b y\,dv + \varepsilon_0\mu_v(J_\delta)$$

with $J_\delta := I_\delta\backslash[a, b]$. μ_v denotes the regular Borel measure on $[0, T]$ associated with the function $\|v(t)\|$. Hence we have

$$\int_a^b w\,dv \leqq \|w\|_Z\mu_v(J_\delta) \quad \forall \delta \in]0, \delta_0].$$

Thus, since $\lim_{\delta\to 0}\mu_v(J_\delta)=0$, we get $\int_a^b w\,dv \leqq 0$. Since w is arbitrary, we conclude that $v|_{[a,b]}=0$. Using the fact that M_{u_0} can be written as the union of such intervals, the result follows. □

For each $u\in U_{ad}$, the set M_u can be written in the form

$$M_u := \bigcup_{i\in I}]a_i, b_i[, \tag{4.9}$$

$$]a_i, b_i[\cap]a_j, b_j[= \varnothing \quad \forall i, j \in I, \qquad i\neq j, \tag{4.10}$$

where I is a finite or countable index set. By $p(\hat{\zeta}, 2(p_2Su_0-y_T)+\zeta_T)$ we denote the solution of (4.4), (4.5). For an optimal control u_0, a functional $\zeta\in Z^*$ is called a multiplier if u_0 and ζ fulfill the equations of Lemma 4.1. With these notations the following assumption can be formulated.

(A) For each optimal control u_0 there exists a multiplier ζ such that

$$p(\hat{\zeta}, 2(p_2Su_0-y_T)+\zeta_T)(b_i-) \neq 0 \quad \forall i \in I.$$

Here, M_{u_0} is decomposed as in (4.9), (4.10). Finally we introduce the set

$$M = \{M_u \mid u \text{ is optimal}\}.$$

THEOREM 4.1. *Let $\nu=0$ and suppose that* (SL) *and* (A) *are fulfilled. Then each optimal control is bang-bang on $M\times\Gamma$, and*

$$|u_0(t,\xi)| = \rho \quad \forall'(t,\xi)\in M\times\Gamma.$$

Moreover, if u_1 is a further optimal control, then u_1 and u_0 coincide almost everywhere on $M\times\Gamma$.

Proof. Let ζ be a multiplier of u_0. Then Lemma 4.3 shows that $v|_{M_{u_0}}=0$ for $v=v_\zeta$, hence, in particular, $v|_{]a_i,b_i[}=0$ for all $i\in\mathbb{N}$. Thus (4.4) implies (with $p = p(\hat{\zeta}, 2(p_2Su_0-y_T)+\zeta_T)$

$$-\frac{\partial p}{\partial t} + Ap = 0 \qquad \text{on }]a_i, b_i[\times\Omega,$$

$$\alpha p|_\Sigma + \beta\frac{\partial p}{\partial n_A} = 0 \qquad \text{on }]a_i, b_i[\times\Gamma.$$

Suppose now that for an $i \in I$ the control u_0 is not bang-bang on $]a_i, b_i[\times \Gamma$. Then by (4.1) there is an open interval $]c, d[\subset]a_i, b_i[$ such that

$$Rp(t, \xi) = 0 \quad \forall (t, \xi) \in]c, d[\times \Gamma$$

(with R as in the proof of Lemma 4.2). Thus by Schmidt and Weck [17, Cor. 2.3] we conclude $p(t, x) = 0$ for all $(t, x) \in]a_i, b_i[\times \Omega$. This implies $p(b_i -) = 0$, which is a contradiction to assumption (A). Hence u_0 is bang-bang on $M_{u_0} \times \Gamma$.

Now let u_1 be a further optimal control. Then we have

$$\tfrac{1}{2} y(u_0)(t) + \tfrac{1}{2} y(u_1)(t) \in \operatorname{int} C \quad \forall t \in M_{u_0} \cup M_{u_1}.$$

This implies $M_{u_0} \cup M_{u_1} \subset M_{u_2}$ if we set $u_2 := \frac{1}{2} u_0 + \frac{1}{2} u_1$. Suppose that there is a set $\Sigma_0 \subset M_{u_1} \times \Gamma$ of positive measure such that $|u_0(t, \xi)| < \rho$ almost everywhere on Σ_0. Then

$$-\rho < u_2(t, \xi) = \tfrac{1}{2} u_0(t, \xi) + \tfrac{1}{2} u_1(t, \xi) < \rho \quad \text{a.e. on } \Sigma_0,$$

which is a contradiction. Hence u_0 is bang-bang on $M_{u_1} \times \Gamma$.

We assume that u_0 is not bang-bang on $M \times \Gamma$. Then there is an interval $]a, b[\subset M$ such that $|u_0(t, \xi)| < \rho$ almost everywhere on $]a, b[\times \Gamma$. Let $d := \frac{1}{2}(a + b)$. There must be an optimal control $\hat{u}$ and an interval $]a_i, b_i[$ with $d \in]a_i, b_i[\subset M_{\hat{u}}$. Hence, as we have seen above, u_0 is bang-bang on $M_{\hat{u}} \times \Gamma$ and in particular on $]d - \delta, d + \delta[\times \Gamma$ (with sufficiently small $\delta > 0$) which is a contradiction. Thus u_0 is bang-bang on $M \times \Gamma$.

If u_1 is a further optimal control, it must coincide with u_0 almost everywhere on $M \times \Gamma$, since both controls are bang-bang on this set. Otherwise, $u_2 := \frac{1}{2} u_0 + \frac{1}{2} u_1$ would be an optimal control not bang-bang on $M \times \Gamma$. □

Remark. If (P) has no state constraint, then (A) is fulfilled if y_T cannot be reached by a feasible control. This is the usual condition needed for the standard bang-bang case (cf. Schmidt and Weck [17]). In the general case, (A) has no such simple interpretation. In some special situations, it is possible to reduce (A) to a condition which is less complicated. Compare in this connection the generalized bang-bang principles in Tröltzsch [18] (see also Mackenroth [10]).

We now turn back to the question of whether a sequence $\{u_h\}$ of discrete optimal controls converges to an optimal control u_0 of (P). Let $N := [0, T] \setminus M$.

THEOREM 4.2. *Let $\nu = 0$ and suppose that the assumptions of Theorem* 2.1 *are fulfilled. Then the following assertions hold with $j = 1, 2$ as in* § 3.

(a) *We have*

$$\lim_{h \to 0} p_2 S_h^j u_h = p_2 S^j u_0.$$

If u_1 is a further optimal control, then $p_2 S u_1 = p_2 S u_0$.

(b) *There is a subsequence $\{w_h\}$ of $\{u_h\}$ which converges weakly in $L^2(\Sigma)$ to u_0. For each such subsequence we have*

$$\lim_{h \to 0} p_1 S_h^j w_h = p_1 S u_0, \quad \text{and} \quad \lim_{h \to 0} p_1 S w_h(t) \in \partial C \quad \forall t \in N.$$

(c) *If in addition* (A) *is true, then*

$$\lim_{h \to 0} \| u_h|_{M \times \Gamma} - u_0|_{M \times \Gamma} \|_{L^q(M \times \Gamma)} = 0 \quad \forall q \in [1, \infty].$$

Proof. Assertions (a) and (b) follow by standard arguments and the compactness of S from Theorem 2.1. Assertion (c) is an immediate consequence of Theorem 4.1 and Mackenroth [11, Lemma 2]. □

Remark. A similar result can be formulated for the one-dimensional case and S_m^3.

The question of the behaviour of $\{u_h\}$ on $N \times \Gamma$ is not answered by the preceding theorem. For $t \in N$ we have $p_1 S u_0(t) \in \partial C$ which often leads to an additional equation

for u_0. Then the second assertion of Theorem 4.2(b) can be interpreted as a result of the defects of the discrete optimal controls on N.

5. Numerical examples. In this section we consider some numerical examples for the case of space dimension $n=1$. A numerical example for $n=2$ can be found in Mackenroth [13].

We set $T=1$, $\Omega=]0, 1[$, $a_{11}=1$, $a_0=0$ and suppose that the boundary conditions are given by

$$\frac{\partial y}{\partial x}(t, 0)=0,$$

$$\alpha y(t, 1)+\frac{\partial y}{\partial x}(t, 1)=u(t),$$

with $\alpha=0.1$. Let $y_T(x)=0.7$, assume that the control constraint is given by

$$0 \leqq u(t) \leqq 1 \quad \forall' t \in [0, T],$$

and define

$$C(t)=\{z \in Z \,|\, (z(t), 1) \leqq \eta(t) \quad \forall t \in [0, T]\}$$

with $\eta \in C[0, T]$. We approximate $y(u)$ by $y_m(u)$ as described in (3.14). This can be easily implemented since for λ_k, v_k well-known explicit formulas are available. Experience has shown that for $m=20$, the Fourier series gives a very good approximation of $y(u)$ and the computational effort remains within reasonable limits.

For the numerical solution, it is of course necessary to discretize also the time t. We do this by using piecewise constant controls and by imposing the state constraint only at finitely many discrete points. Then it is an easy task to convert the discrete problems into a quadratic programming problem with inequality constraints. For the numerical solution of these problems we have used the program SOL/QPSOL by Gill et al. which worked very well. This program is available in the NAG-Library.

In our first example we consider the bang-bang case. If no state constraint is present, the optimal control is given by

$$u_0(t)=\begin{cases}1 & \text{if } t \in [0, t_s] \\ 0 & \text{if } t \in]t_s, 1]\end{cases}$$

with $t_s=0.767$. The integrated state, i.e., the function $(y(u)(\cdot), 1)$ is shown in Fig. 1. This picture shows that a state constraint with

$$\eta(t)=\begin{cases}-t+1 & \text{if } t \in [0, 0.6], \\ 2.5t-1.1 & \text{if } t \in [0.6, 1],\end{cases}$$

must affect the solution. The corresponding optimal control and integrated state (obtained by numerical computations) are depicted in Figs. 2 and 3.

We see that in this case, N consists only of a single point t_N. The optimal control u_0 is bang-bang on the whole interval $[0, 1]$ and has a jump at t_N. Thus u_0 behaves as predicted by the theory.

If we choose a differentiable function for η the situation becomes different. For

$$\eta(t)=0.1e^{3t-0.6}+0.15 \quad \forall t \in [0, 1]$$

the results are shown in Figs. 4 and 5. In this case, N consists of an interval with nonempty interior. On this set, u_0 seems to be smooth.

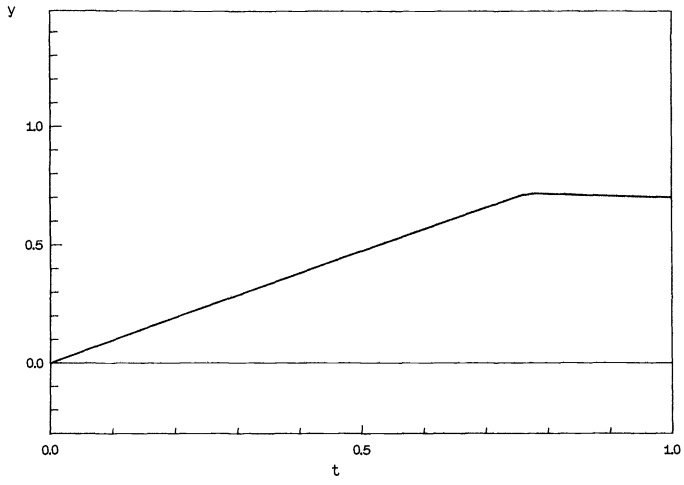

FIG. 1

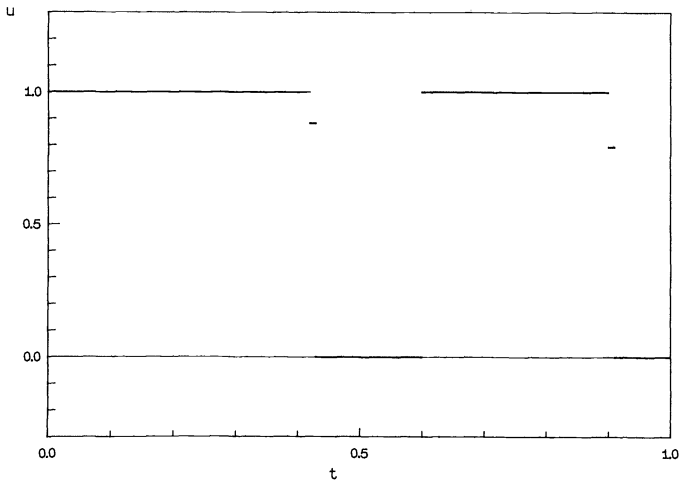

FIG. 2

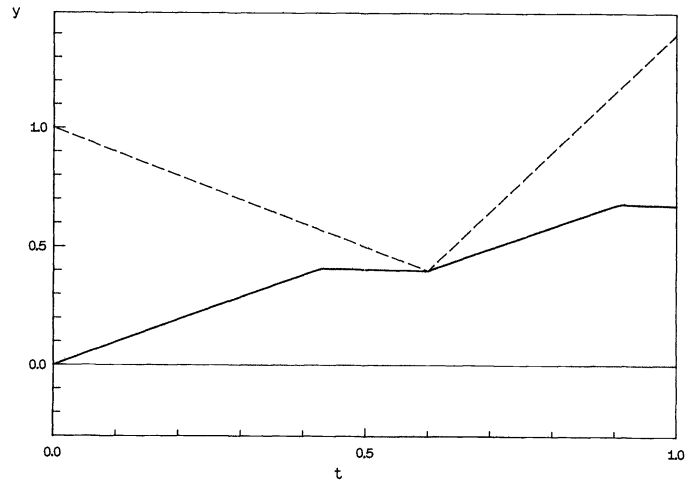

FIG. 3

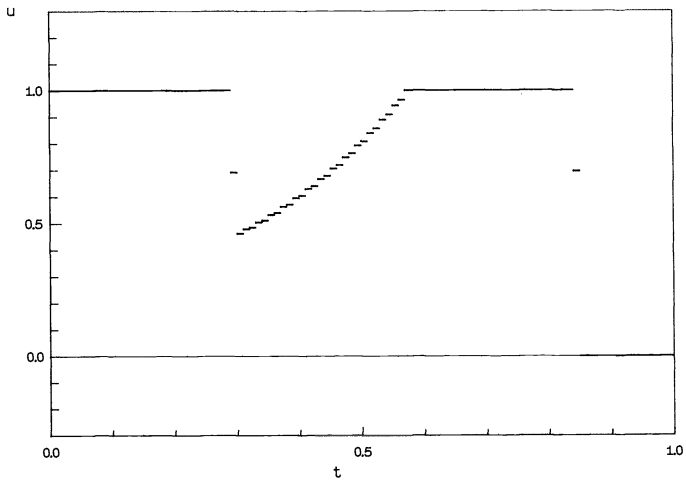

FIG. 4

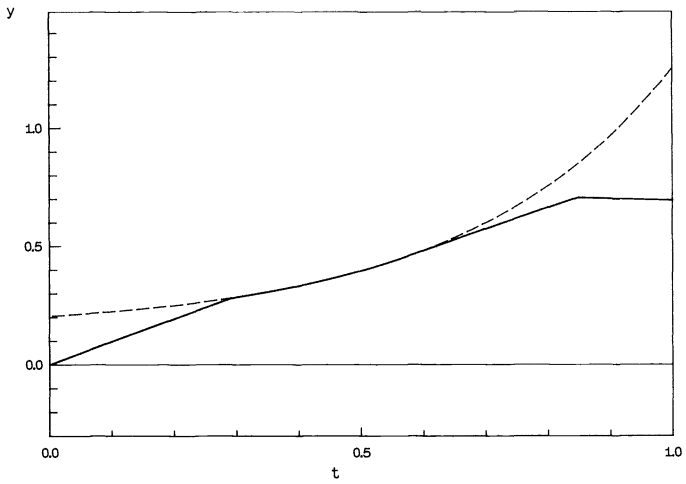

FIG. 5

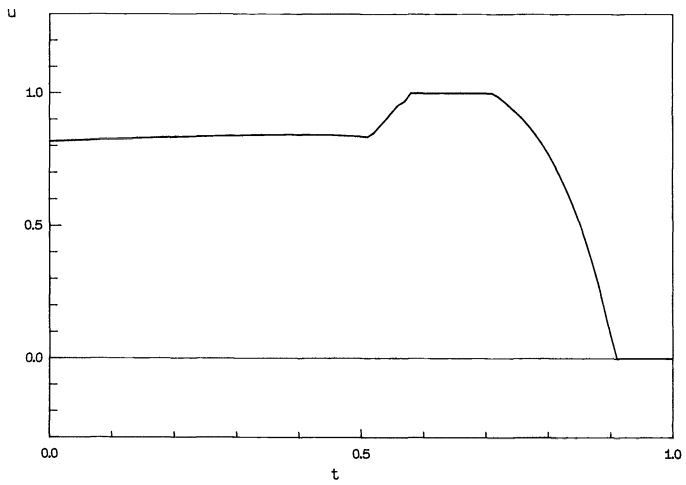

FIG. 6

Now let $\nu = 0.01$ and all other data as in the previous example. The corresponding optimal control is shown in Fig. 6. Here piecewise linear functions are chosen for the approximation of the controls. Further numerical examples with some additional theoretical results can be found in Alt and Mackenroth [1].

All computations have been performed on a VAX 11/780 in double precision. A typical computation time was about 15 seconds.

Acknowledgment. The authors thank one of the referees for a helpful hint concerning the proof of Proposition 3.1.

REFERENCES

[1] W. ALT AND U. MACKENROTH, *On the numerical solution of state constrained coercive optimal control problems*, in Optimal Control of Partial Differential Equations, K. H. Hoffmann and W. Krabs, eds., Proceedings of a conference held at Oberwolfach, December 1982, Birkhäuser-Verlag, 1984.

[2] V. BARBU and T. PRECUPANU, *Convexity and optimization in Banach spaces*, D. Reidel Publishing Company, Dordrecht-Boston-Lancaster, 1986.

[3] N. DINCULEANU, *Vector Measures*, Pergamon Press, Oxford, VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.

[4] H. O. FATTORINI, *Optimal control of nonlinear systems: convergence of suboptimal controls* I, in Proceedings of the Special Session on Operator Methods in Optimal Control Problems, Annual AMS Meeting, New Orleans, January 1986.

[5] G. KNOWLES, *Finite element approximations of parabolic time optimal control problems*, SIAM J. Control Optim., 30 (1982), pp. 414-427.

[6] I. LASIECKA, *State constraint control problems for parabolic systems: regularity of optimal solutions*, Appl. Math. Optim., 6 (1980), pp. 1-29.

[7] ———, *Boundary control of parabolic systems: finite element approximation*, Appl. Math. Optim., 6 (1980), pp. 31-62.

[8] ———, *Ritz-Galerkin approximation of the time optimal boundary control problem for parabolic systems with Dirichlet boundary conditions*, SIAM J. Control Optim., 22 (1984), pp. 477-500.

[9] U. MACKENROTH, *Optimalitätsbedingungen und Dualität bei zustandsrestringierten parabolischen Kontrollproblemen*, Math. Operationsforsch. Statist. Sér. Optim., 12 (1981), pp. 65-89.

[10] ———, *Bang-bang controls for time optimal parabolic boundary control problems with integral state constraints*, in Optimization: Theory and Algorithms, J. B. Hiriart-Urruty, W. Oettli, and J. Stoer, eds., Proceedings of a conference held at Confolant (France), March 1981, Lecture Notes in Pure and Applied Mathematics, 86 (1983), pp. 213-223.

[11] ———, *Some remarks on the numerical solution of bang-bang type optimal control problems*, Numer. Funct. Anal. Optim., 5 (1983), pp. 467-484.

[12] ———, *On parabolic distributed control problems with restrictions on the gradient*, Appl. Math. Optim., 12 (1983), pp. 69-95.

[13] ———, *Numerical solution of some parabolic boundary control problems by finite elements*, in Control Problems for Systems Described by Partial Differential Equations and Applications, I. Lasiecka and R. Triggiani, eds., Proceedings of the IFIP-WG 7.2 Working Conference held at the University of Florida, Gainesville, 1986, Lecture Notes in Control and Information Sciences 97, 1987, pp. 325-335.

[14] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control constraint optimal control problems*, Appl. Math. Optim., 8 (1981), pp. 69-85.

[15] S. M. ROBINSON, *Regularity and stability for convex multivalued functions*, Math. of Oper. Res., 1 (1976), pp. 130-143.

[16] D. WASHBURN, *A bound on the boundary input map for parabolic equations with applications to time optimal control*, SIAM J. Control Optim., 17 (1979), pp. 652-671.

[17] E. P. J. G. SCHMIDT AND N. WECK, *On the boundary behaviour of solutions to elliptic and parabolic equations with applications to boundary control for parabolic equations*, SIAM J. Control Optim., 16 (1978), pp. 593-598.

[18] F. TRÖLTZSCH, *Optimality conditions for parabolic control problems and applications*, Teubner, Leipzig, 1984.

[19] ———, *Semidiscrete finite element approximation of parabolic boundary control problems—convergence of switching points*, ISNM, 78 (1987), pp. 219-232.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 737-757, July 1989

004

AN ELLIPSOID TRUST REGION BUNDLE METHOD FOR NONSMOOTH CONVEX MINIMIZATION*

KRZYSZTOF C. KIWIEL†

Abstract. This paper presents a bundle method of descent for minimizing a convex (possibly nonsmooth) function f of several variables. At each iteration the algorithm finds a trial point by minimizing a polyhedral model of f subject to an ellipsoid trust region constraint. The quadratic matrix of the constraint, which is updated as in the ellipsoid method, is intended to serve as a generalized "Hessian" to account for "second-order" effects, thus enabling faster convergence. The interpretation of generalized Hessians is largely heuristic, since so far this notion has been made precise by J. L. Goffin only in the solution of linear inequalities. Global convergence of the method is established and numerical results are given.

Key words. nonsmooth optimization, nondifferentiable programming, convex programming, descent methods, ellipsoid algorithm

AMS(MOS) subject classifications. primary 65K05; secondary 90C25

1. Introduction. This paper presents a readily implementable algorithm for minimizing a convex (possibly nonsmooth) real-valued function f defined on R^N. We suppose that the set of minimum points of f:

$$X^* = \operatorname{Arg\,min} f = \{x^* \in R^N : f(x^*) \leqq f(x) \forall x \in R^N\}$$

is nonempty, and that we know center $x_c \in R^N$ and radius $r > 0$ of some ball $E = \{x \in R^N : |x - x_c| \leqq r\}$ that intersects X^*. The algorithm requires only the computation of $f(x)$ and one arbitrary subgradient $g_f(x) \in \partial f(x)$ of f at each $x \in R^N$.

The performance of the several existing methods for minimizing f depends on the shape of the level sets of f. The bundle methods of descent (see, e.g., [K2] for their survey), which can be derived by introducing a regularizing quadratic term in the cutting plane method [K1], seem to perform best when f is close to being piecewise linear (polyhedral) and the Haar condition [H1] holds at the minimum point. These methods are sensitive to objective scaling (multiplication of f by a positive number), since so far no scale-invariant rules are known for choosing the weights of their quadratic terms. On the other hand, the ellipsoid method (see [S1] and [Y1]), which can be described as a variable metric subgradient optimization method [G5], seems to work well when f has very elongated level sets, whereas its performance deteriorates for polyhedral functions with "fat" level sets (see [Y2, § 9.5] and [G5]). This method is insensitive to objective scaling.

The algorithm of this paper attempts to combine the best features of the ellipsoid and bundle methods. At each iteration it finds a trial point by minimizing a polyhedral model of f subject to an ellipsoid trust region constraint. The quadratic matrix of the constraint, which is updated as in the ellipsoid method, is intended to serve as a generalized "Hessian" to account for "second-order" effects, with the purpose of enabling faster convergence when the Haar condition does not hold. The subproblem of finding the trial point is solved approximately by estimating the Lagrange multiplier of its constraint and solving the resulting quadratic programming subproblem. In effect, the algorithm may also be viewed as a bundle method with an automatic choice of the quadratic term and its weight, which is, in principle, insensitive to objective scaling.

* Received by the editors July 15, 1987; accepted for publication (in revised form) May 25, 1988. This research was supported by The Polish Academy of Sciences, Project CPBP.02.15.

† Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447 Warsaw, Poland.

Our algorithm is intended for complex functions of relatively few variables, for which one function and subgradient evaluation dominates the effort per iteration involved in variable metric updates and quadratic programming subproblems. Hence it tries to minimize the number of function evaluations required to achieve a given accuracy. In particular, it evaluates f not at the center of the current ellipsoid, as do the existing ellipsoid methods (see [S2], [E1], [E2]), but at a point that should have a lower objective value according to the accumulated model of f.

Our interpretation of the generated ellipsoids as generalized "Hessians" is largely heuristic, since so far this notion has been made precise by Goffin [G4], [G7] only in the solution of linear inequalities. Alternative approaches to incorporating "second-order" models in the bundle methods are given in [L3], [L4], and [M1]. So far, they have not produced implementable algorithms.

We prove that the sequence of points generated by the method minimizes f. Rate-of-convergence results are still missing, but we report some encouraging numerical experience. Also, we will show that there is much freedom in implementing the algorithm, and we have only begun to explore some of the possibilities. In particular, we have been using simple ellipsoid cuts, whereas more refined ellipsoid updates would probably be more efficient.

We refer the reader to [A1] and [B1] for surveys and bibliographies of the ellipsoid method, and to [S3] for some of its modifications.

The paper is organized as follows. In § 2 we derive the algorithm, which is stated in detail in § 3. Its convergence is established in § 4. Various more efficient ellipsoid updating strategies are discussed in § 5. Section 6 describes an implementation of the method. Subgradient aggregation is introduced in § 7. Numerical results are reported in § 8. Section 9 concludes the paper.

We use the following notation. We denote by $\langle \cdot, \cdot \rangle$ and $|\cdot|$, respectively, the usual inner product and norm in finite-dimensional, real Euclidean space R^N. We use x_i to denote the ith component of the vector x. Superscripts are used to denote different vectors, e.g., x^1 and x^2. All vectors are column vectors. However, for convenience we sometimes write (x, y) for $(x^T, y^T)^T$, where T denotes transposition. For an $N \times N$ symmetric positive definite matrix A, we let A^{-1} denote the inverse of A, $A^{-T} = (A^{-1})^T$, $\langle x, y \rangle_A = \langle Ax, y \rangle = x^T Ay$, $|x|_A = \langle Ax, x \rangle^{1/2}$. The volume Vol (S) of a bounded measurable set S in R^N is its N-dimensional Lebesgue measure.

For any $x \in R^N$,

$$\partial f(x) = \{g \in R^N : f(y) \geqq f(x) + \langle g, y - x \rangle \forall y \in R^N\}$$

denotes the subdifferential of f at x. The mapping $\partial f(\cdot)$ is locally bounded and f is continuous (see, e.g., [D1, § 1.7.1 and Thm. 1.4.1]). $T(\alpha) = \{x \in R^N : f(x) \leqq \alpha\}$ is the α-level set of f.

2. Derivation of the method. The algorithm to be described will generate two sequences of points $\{x^k\}_{k=1}^{\infty}$ and $\{y^k\}_{k=1}^{\infty}$ in R^N, where $x^1 = y^1$ is a given starting point. The sequence $\{x^k\}$ will satisfy $f(x^{k+1}) < f(x^k)$ if $x^{k+1} \neq x^k$, and $f(x^k) \downarrow \min f$. The auxiliary trial points y^k will be used for computing $f(y^k)$ and $g^k = g_f(y^k)$ for all k (with $y^{k+1} = x^{k+1}$ if $x^{k+1} \neq x^k$; see below). Also a sequence of ellipsoids

$$E_k = \{x \in R^N : |x - x_c^k|_{A_k} \leqq 1\}$$

with centers $x_c^k \in R^n$ and symmetric positive definite matrices A_k will be generated such that $E_k \cap X^* \neq \varnothing$ for all k, where E_1 is a given starting ellipsoid.

At the kth iteration, the algorithm will try to find a point y^{k+1} such that $f(y^{k+1}) < f(x^k)$. Ideally, y^{k+1} should minimize f. Since $E_k \cap X^* \neq \varnothing$, we may restrict the minimiz-

ation to E_k, i.e., we may consider the subproblem

$$\text{minimize } f(y) \quad \text{over all } y \in E_k. \tag{2.1}$$

The trial point y^{k+1} will solve an approximate but manageable version of subproblem (2.1), which is derived as follows.

The algorithm will use the following polyhedral approximation to f:

$$\hat{f}^k(x) = \max\{f_j(x): j \in J^k\} \quad \text{for all } x,$$

where J^k is a subset of $\{1, \cdots, k\}$ and

$$f_j(x) = f(y^j) + \langle g_f(y^j), x - y^j \rangle \quad \text{for all } x$$

is the jth linearization of f satisfying $f(x) \geqq f_j(x)$ for all x. Replacing f by $\hat{f}^k$ in (2.1), we obtain the subproblem

$$\text{minimize } \hat{f}^k(y) \quad \text{over all } y \in E_k$$

and its equivalent form

$$\begin{aligned} &\text{minimize } u \text{ over all } (y, u) \in R^N \times R \\ &\text{satisfying } f_j(y) \leqq u \quad \text{for all } j \in J^k, \\ &\tfrac{1}{2}|y - x_c^k|_{A_k}^2 \leqq \tfrac{1}{2}. \end{aligned} \tag{2.2}$$

Let $\hat{\eta}^k$ denote the optimal Lagrange multiplier for the quadratic constraint of (2.2). It is not easy—and apparently not advisable—to solve (2.2) too accurately; the algorithm will find (y^{k+1}, u^k) to

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\eta^k |y - x_c^k|_{A_k}^2 + u \quad \text{over all } (y, u) \in R^{N+1} \\ &\text{satisfying } f_j(y) \leqq u \quad \text{for all } j \in J^k \end{aligned} \tag{2.3}$$

for some $\eta^k \geqq \hat{\eta}^k$, so that $y^{k+1} \in E_k$. The search for η^k, which will test increasing values of η^k, will be similar to that employed in the trust region methods (see [M2] for a survey). Section 6 will suggest another argument in favor of larger values of η^k.

Another motivation for subproblem (2.3) stems from the fact that, in a suitably transformed space, it reduces to the subproblems of the bundle methods of [K2, Chap. 2], thus inheriting their useful theoretical and computational properties. More specifically, the direction $\bar{d}^k = y^{k+1} - x_c^k$ and the predicted objective decrease

$$v^k = \hat{f}^k(y^{k+1}) - f(x^k) = u^k - f(x^k) \tag{2.4}$$

can be found by solving the subproblem

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\eta_k |\bar{d}|_{A_k}^2 + v \quad \text{over all } (\bar{d}, v) \in R^{N+1} \\ &\text{satisfying } -\bar{\alpha}_j^k + \langle g^j, \bar{d} \rangle \leqq v \quad \text{for } j \in J^k, \end{aligned} \tag{2.5}$$

where

$$\bar{\alpha}_j^k = f(x^k) - f_j(x_c^k) \quad \text{for } j \in J^k. \tag{2.6}$$

Let us represent the symmetric and positive definite matrices A_k and $B_k = A_k^{-1}$ as

$$A_k = \tilde{A}_k^T \tilde{A}_k \quad \text{and} \quad B_k = \tilde{B}_k \tilde{B}_k^T,$$

where $\tilde{B}_k = \tilde{A}_k^{-1}$ is a nonsingular $N \times N$ matrix (e.g., $\tilde{B}_k^T$ is the Cholesky factor B_k). Consider the space transformation

$$x \to \tilde{x} = \tilde{A}_k x,$$

which maps x^k, x_c^k, y^j, y^{k+1}, and $\bar{d}^k$ into $\tilde{x}^k$, $\tilde{x}_c^k$, $\tilde{y}^j$, $\tilde{y}^{k+1}$, and $\tilde{d}^k$, respectively. In the transformed space ($\tilde{x}$-space), $\tilde{g}^j = \tilde{B}_k^T g^j$ are the subgradients of the function $\tilde{f}(\tilde{x}) = f(\tilde{A}_k^{-1}\tilde{x})$ at $\tilde{y}^j$, for $j \in J^k$, whereas $(\tilde{d}^k, v^k)$ solves the transformed subproblem

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}\eta^k|\tilde{d}|^2 + v \quad \text{over all } (\tilde{d}, v) \in R^{N+1} \\ &\text{satisfying } -\bar{\alpha}_j^k + \langle \tilde{g}^j, \tilde{d}\rangle \leqq v \quad \text{for } j \in J^k. \end{aligned} \tag{2.7}$$

The bundle methods of [K2, Chap. 2] use subproblem (2.7) for generating a descent direction for $\tilde{f}$ at $\tilde{x}_c^k$. (In fact, they use $\tilde{x}_c^k = \tilde{x}^k$ and $\eta^k = 1$ in (2.7), but η^k can be suppressed in (2.7) by replacing $\tilde{A}_k$ with $(\eta^k)^{1/2}\tilde{A}_k$ in the transformation.) Therefore, by transforming "back" the results of [K2, pp. 49, 64], we can establish the following properties of subproblem (2.5). Let $\lambda_j^k, j \in J^k$, denote the (possibly nonunique) Lagrange multipliers of (2.5), and let

$$(p^k, \bar{\alpha}_p^k) \equiv \sum_{j \in J^k} \lambda_j^k (g^j, \bar{\alpha}_j^k) \tag{2.8}$$

and $\tilde{p}^k = \tilde{B}_k^T p^k$. Then $\tilde{d}^k = -\tilde{p}^k/\eta^k$, $-v^k = |\tilde{p}^k|^2/\eta^k + \bar{\alpha}_p^k$,

$$\bar{d}^k = -\frac{1}{\eta^k} B_k p^k, \tag{2.9a}$$

$$v^k = -\left\{\frac{1}{\eta^k}|p^k|_{B_k}^2 + \bar{\alpha}_p^k\right\}. \tag{2.9b}$$

Moreover,

$$w^k = \frac{1}{2\eta^k}|p^k|_{B_k}^2 + \bar{\alpha}_p^k \tag{2.10}$$

is the optimal value of the dual subproblem

$$\begin{aligned} &\text{minimize } \frac{1}{2\eta^k}\left|\sum_{j \in J^k} \lambda_j \tilde{g}^j\right|^2 + \sum_{j \in J^k} \lambda_j \bar{\alpha}_j^k, \\ &\text{subject to } \sum_{j \in J} \lambda_j = 1, \quad \lambda_j \geqq 0 \quad \text{for } j \in J^k, \end{aligned} \tag{2.11}$$

whose solution set coincides with the set of Lagrange multipliers of (2.5). In particular,

$$\sum_{j \in J^k} \lambda_j^k = 1, \quad \lambda_j^k \geqq 0 \quad \text{for } j \in J^k. \tag{2.12}$$

In general, we would like the ellipsoid $E = E_k$ to be a tight approximation to (a portion of) X^*, since then the point $y^{k+1} \in E_k$ would be close to optimal. Whenever we identify a redundant portion E^- of E such that $E^- \cap X^* = \varnothing$, we may reduce E by replacing it with a smaller ellipsoid E_+ that contains the remaining portion $E \setminus E^-$ as in the ellipsoid methods (see, e.g., [S2]). The ellipsoid methods update E as follows. Using one or more cutting planes, they choose a portion of E that contains $E \cap X^*$ and no more than half of E, and let E_+ be the smallest-volume ellipsoid containing this portion. To ensure that at least half of E is cut off, so that $\text{Vol}(E_+) \leqq q\ \text{Vol}(E)$ with $q < e^{-1/(2N+2)} < 1$ (see [T1]), they use the hyperplane $\{x \in R^N : \langle g_f(x_c^k), x - x_c^k\rangle = 0\}$, or its translation (if $g_f(x_c^k) = 0$, they terminate with $x_c^k \in X^*$). Our algorithm will use the bounding hyperplane of

$$H_p^k = \{x \in R^N : \langle p^k, x - x_c^k\rangle \leqq \bar{\alpha}_p^k\} \tag{2.13}$$

defined via (2.8). This will save the computation of $g_f(x_c^k)$, but will provide a significant volume reduction only when $\bar{\alpha}_c^k \leqq 0$. More specifically, if we let E_+ be the least volume ellipsoid containing $E \cap H_p^k$, then $\mathrm{Vol}\,(E_+) \leqq q\,\mathrm{Vol}\,(E)$ if $\bar{\alpha}_p^k \leqq 0$ and $p^k \neq 0$, whereas $E_+ \cap X^* \neq \varnothing$ if $(E \cap H_p^k) \cap X^* \neq \varnothing$, which will hold if

$$T(f(x^k)) \subset H_p^k, \tag{2.14}$$

because $E_k \cap X^* \neq \varnothing$. Now, multiplying by λ_j^k the relations

$$\begin{aligned} f(x) &\geqq f(y^j) + \langle g_f(y^j),\, x - y^j\rangle = f_j(x) \\ &= f(x^k) + \langle g^j,\, x - x_c^k\rangle - \bar{\alpha}_j^k \quad \text{for all } x \end{aligned} \tag{2.15}$$

and summing over $j \in J^k$, we deduce from (2.8) and (2.12) that

$$f(x) \geqq f(x^k) + \langle p^k,\, x - x_c^k\rangle - \bar{\alpha}_p^k \quad \text{for all } x, \tag{2.16}$$

which establishes (2.14). Moreover, by (2.16), the algorithm may stop with $x^k \in X^*$ if $\bar{\alpha}_p^k \leqq 0$ and $p^k = 0$. Thus we may replace E with E_+ to obtain the desired volume reduction if $\bar{\alpha}_p^k \leqq 0$.

In geometric terms, the *aggregate cut* discussed above is provided by the *aggregate linearization*

$$\tilde{f}^k(x) = \sum_{j \in J^k} \lambda_j^k f_j(x) = \tilde{f}^k(x_c^k) + \langle p^k,\, x - x_c^k\rangle \quad \text{for all } x$$

with $\tilde{f}^k(x_c^k) = f(x^k) - \bar{\alpha}_p^k$; relation (2.16) means that $f(x) \geqq \tilde{f}^k(x)$ for all x, and we have $H_p^k = \{x \in R^N : \tilde{f}^k(x) \leqq f(x^k)\}$. A *deep* cut with $x_c^k \notin H_p^k$ is obtained if $\tilde{f}^k(x_c^k) > f(x^k)\,(\bar{\alpha}_p^k < 0)$. On the other hand, relation (2.15) shows that the "ordinary" linearizations f_j define cuts (called *supercuts* in [S2]) based on the relations

$$H_j^k = \{x \in R^n : \langle g^j,\, x - x_c^k\rangle \leqq \bar{\alpha}_j^k\} = \{x : f_j(x) \leqq f(x^k)\}, \qquad T(f(x^k)) \subset H_j^k. \tag{2.17}$$

Thus our aggregate cut is a convex combination of "ordinary" cuts (cf. (2.8) and (2.12)); it reduces to the *surrogate cut* of [G8] when f is polyhedral and $f(x^k) = \min f$. More will be said about cuts in § 5.

A useful stopping criterion can be derived from (2.16) as follows. Since $B_k = A_k^{-1}$ and the Cauchy-Schwarz inequality yields

$$\begin{aligned} |\langle p^k,\, x - x_c^k\rangle| &= |\langle B_k p^k,\, x - x_c^k\rangle_{A_k}| \leqq |B_k p^k|_{A_k} |x - x_c^k|_{A_k} \\ &= |p^k|_{B_k} |x - x_c^k|_{A_k}, \end{aligned}$$

we may set $x = x^* \in E_k \cap X^*$ in (2.16) to obtain

$$f(x^k) \leqq \min f + |p^k|_{B_k} + \bar{\alpha}_p^k. \tag{2.18}$$

Thus the algorithm may stop if $|p^k|_{B_k} + \bar{\alpha}_p^k\,(\geqq 0)$ is sufficiently small.

If neither an ellipsoid update nor termination occur, i.e., $\bar{\alpha}_p^k > 0$ and $p^k \neq 0$, then the predicted objective decrease v^k is negative (cf. (2.4) and (2.9b)). To ensure a significant objective reduction, the algorithm will take a *serious step* from x^k to $x^{k+1} = y^{k+1}$ only if

$$f(y^{k+1}) \leqq f(x^k) + m v^k, \tag{2.19}$$

where $m \in (0, 1)$ is a parameter. Otherwise, a *null step* with $x^{k+1} = x^k$ will occur, but the new linearization f_{k+1} of f at y^{k+1} will contribute to our finding a better next trial point y^{k+2}.

3. The method. We shall now state the simplest version of the method, postponing more efficient modifications until §§ 5 and 6.

ALGORITHM 3.1.

Step 0 (Initialization). Choose a point $x_c^1 \in R^N$ and a symmetric positive definite $N \times N$ matrix A_1 such that the ellipsoid $E_1 = \{x \in R^N : |x - x_c^1|_{A_1} \leqq 1\}$ satisfies $E_1 \cap X^* \neq \emptyset$. Select a starting point $x^1 \in R^N$, a final accuracy tolerance $\varepsilon_s \geqq 0$, a line search parameter $m \in (0, 1)$ and weight updating parameters $\eta_u > 0$ and $\chi \in (1, 100]$. Set $y^1 = x^1$ and $J^1 = \{1\}$. Compute $f(y^1)$, $g^1 = g_f(y^1)$ and $f_1(x_c^1) = f(y^1) + \langle g^1, x_c^1 - y^1 \rangle$. Choose $\eta^1 \in (0, \eta_u]$. Set the iteration counter $k = 1$ and the counter of serious steps $L = 0$. Set $k(0) = 1$ ($k(L)$ will denote the iteration number of the Lth (latest) serious step).

Step 1 (Direction finding). Find the solution $(\bar{d}^k, v^k)$ and Lagrange multipliers λ_j^k, $j \in J^k$, of subproblem (2.5), with $\bar{\alpha}_j^k$, $j \in J^k$, given by (2.6). Compute p^k and $\bar{\alpha}_p^k$ by (2.8).

Step 2 (Stopping criterion). If $|p^k|_{B_k} + \bar{\alpha}_p^k \leqq \varepsilon_s$, terminate. If $\bar{\alpha}_p^k > 0$, go to Step 4; otherwise, continue.

Step 3 (Ellipsoid updating). Find $x_c^+ \in R^N$ and a symmetric positive definite $N \times N$ matrix A_+ such that $E_+ = \{x \in R^N : |x - x_c^+|_{A_+} \leqq 1\}$ is the smallest-volume ellipsoid containing $E_k \cap H_p^k$, where H_p^k is given by (2.13). Replace x_c^k and A_k by x_c^+ and A_+, respectively, choose $\eta^+ \in (0, \eta_u]$, replace η^k by η^+, and go to Step 1.

Step 4 (Weight updating). If $|\bar{d}^k|_{A_k} \leqq 1$, go to Step 5. Otherwise, choose $\eta^+ \in [\chi\eta^k, 100\eta^k]$, set $\eta^k = \eta^+$, and go to Step 1.

Step 5 (Line search). Set $y^{k+1} = x_c^k + \bar{d}^k$ and compute $f(y^{k+1})$ and $g^{k+1} = g_f(y^{k+1})$. If $f(y^{k+1}) \leqq f(x^k) + mv^k$, set $x^{k+1} = y^{k+1}$, $k(L+1) = k+1$ and increase the counter of serious steps L by 1. Otherwise, set $x^{k+1} = x^k$.

Step 6 (Subgradient selection). Set $\hat{J}^k = \{j \in J^k : \lambda_j^k \neq 0\}$ and choose a set J^{k+1} satisfying $\hat{J}^k \cup \{k+1\} \subset J^{k+1} \subset J^k \cup \{k+1\}$.

Step 7. If $x^{k+1} = x^k$, set $\eta^{k+1} = \eta^k$; otherwise, choose $\eta^{k+1} \in (0, \eta_u]$. Set $x_c^{k+1} = x_c^k$ and $A_{k+1} = A_k$. Increase k by 1 and go to Step 1.

A few comments on the algorithm are in order.

Guidelines for choosing an initial ellipsoid can be found, for instance, in [G8] and [E2]. The obvious choice is to let $x_c^1 = x^1$ and $A_1 = (1/r)I$, where I is the identity matrix and $r > 0$ estimates the Euclidean distance from x^1 to X^*. It is reassuring to know that even if we had $E_1 \cap X^* = \emptyset$, the algorithm would still minimize f on E_1, as will be proved in § 4.

Step 1 can be implemented with the quadratic programming routine of [K3] (see § 6).

Termination at Step 2 implies that $f(x^k) \leqq \min f + \varepsilon_s$ (cf. (2.18)). (This estimate would be weakened to $f(x^k) \leqq \min \{f(x) : x \in E_1\} + \varepsilon_s$ if we had $E_1 \cap X^* = \emptyset$; see § 4.)

The ellipsoid update at Step 3 is well defined, since the algebraic distance (in the metric defined by $|\cdot|_{A_k}$) from x_c^k to H_p^k

$$\omega_p^k = -\bar{\alpha}_p^k / |p^k|_{B_k} = -\bar{\alpha}_p^k / |\tilde{p}^k| \tag{3.1}$$

satisfies $\omega_p^k \in [0, 1)$ when $\bar{\alpha}_p^k \leqq 0$ and $|\tilde{p}^k| + \bar{\alpha}_p^k > 0$ after Step 2; note that in the $\tilde{x}$-space $\tilde{E}_k = \tilde{A}_k E_k$ is the unit ball, whereas ω_p^k is the distance from $\tilde{x}_c^k$ to $\tilde{H}_p^k = \tilde{A}_k H_p^k = \{\tilde{x} : \langle \tilde{p}^k, \tilde{x} - \tilde{x}_c^k \rangle \leqq \bar{\alpha}_p^k\}$. Hence we can compute x_c^+ and $B_+ = A_+^{-1}$ as in [G8]; see § 6. At the kth iteration, an infinite number of returns to Step 1 from Steps 3 and 4 is possible only when $x^k \in X^*$ (see § 4), an unlikely situation when f is not polyhedral.

The restrictions on the choice of the weighting coefficient η^k attempt to limit its growth. Specific choices of η^k will be discussed in § 6. Here we may observe that if there is a cycle between Steps 1 and 4 without ellipsoid updates and we choose $\eta^+ = \chi\eta^k$ at Step 4, then the cycle will terminate (see § 4) with $\eta^k/\chi < \hat{\eta}^k \leqq \eta^k$.

Observe that $x^k \in E_k$ for all k if $x^1 \in E_1$, since x^k is never cut off at Step 3 due to (2.14), whereas $y^{k+1} \in E_k$ by construction, for all k.

The quadratic programming routine of [K3] will compute at most $N+1$ nonzero multipliers λ_j^k. Hence at Step 6 we can choose a set J^{k+1} with at most $N+2$ elements. This number of stored subgradients may be reduced by using subgradient aggregation (§ 7).

4. Convergence. In this section we show that the algorithm minimizes f. Naturally, we assume that the final accuracy tolerance ε_s is set to zero (and that $E_1 \cap X^* \neq \emptyset$ at Step 0).

We start with the ellipsoid updates.

LEMMA 4.1. *If Algorithm* 3.1 *did not stop before the kth iteration, then at Step* 1 *we have*

$$T(f(x^k)) \cap E_1 \subset E_k. \tag{4.1}$$

Proof. If $E_k = E_1$, the inclusion is obvious. Hence, suppose that for some $j < k$ we have $E_1 \cap T(f(x^j)) \subset E_j$ at Step 1 and a new ellipsoid E_+ is constructed at Step 3. Then $E_1 \cap T(f(x^j)) \subset E_j \cap T(f(x^j)) \subset E_j \cap H_p^j \subset E_+$ (cf. (2.14)). Since $f(x^{j+1}) \leqq f(x^j)$, the desired conclusion follows by induction.

Since $E_1 \cap X^* \neq \emptyset$ by assumption, relation (4.1) implies that $E_k \cap X^* \neq \emptyset$. Hence we may use (2.18) to obtain Lemma 4.2.

LEMMA 4.2. *If Algorithm* 3.1 *terminates at the kth iteration, then* $x^k \in X^*$.

From now on we suppose that the algorithm does not terminate.

Due to Lemma 4.1, the case of an infinite number of ellipsoid updates may be analyzed as in [G6].

LEMMA 4.3. *If Algorithm* 3.1 *executes Step* 3 *infinitely many times, then either* k *stays bounded and* $x^k \in X^*$, *or* $f(x^k) \downarrow \min f$ *as* $k \to \infty$.

Proof. By construction, when Step 3 is entered with $\omega_p^k \in [0, 1)$ (see (3.1)), E_+ satisfies $\operatorname{Vol}(E_+) < e^{-1/(2N+2)} \operatorname{Vol}(E_k)$ (see [T1]), and E_+ becomes E_k until the next update. Hence infinitely many updates lead to $\operatorname{Vol}(E_k) \downarrow 0$. Now, to derive a contradiction, suppose that there are infinitely many iterations with $f(x^k) \geqq \bar{f}$, where $\bar{f} > f(x^*)$ for some $x^* \in E_1 \cap X^*$. Then $\operatorname{Vol}(T(\bar{f}) \cap E_1) > 0$ from the continuity of f, whereas (4.1) yields $T(\bar{f}) \cap E_1 \subset E_k$. Thus, $0 < \operatorname{Vol}(T(\bar{f}) \cap E_1) \leqq \operatorname{Vol}(E_k)$ for all k, a contradiction to $\operatorname{Vol}(E_k) \downarrow 0$. Since $\{f(x^k)\}$ is nonincreasing, we deduce that $f(x^k) \downarrow f(x^*)$ if $k \to \infty$. If Step 3 is executed infinitely many times for some fixed k, the same arguments show that $x^k \in X^*$.

From now on we suppose that only a finite number of ellipsoid updates occur. Then there exist $k_E \geqq 1$ and an ellipsoid $\bar{E} = \{x : |x - \bar{x}_c|_{\bar{A}} \leqq 1\}$ with center $\bar{x}_c$ and a symmetric positive definite matrix $\bar{A}$ such that after the last return from Step 3 to Step 1 that occurred at iteration k_E, if any, Step 1 is entered with

$$E_k = \bar{E}, \quad x_c^k = \bar{x}_c, \quad A_k = \bar{A} \quad \text{if } k \geqq k_E. \tag{4.2}$$

Let $\gamma > 0$ denote the square root of the minimum eigenvalue of $\bar{A}$, so that

$$\gamma |x| \leqq |x|_{\bar{A}} \quad \text{for all } x. \tag{4.3}$$

We may now show that the algorithm cannot cycle infinitely between Steps 1 and 4 when its ellipsoid stays constant.

LEMMA 4.4. *Under the preceding assumptions, Algorithm* 3.1 *executes Step* 5 *at each iteration and there exists* $\hat{\eta} > 0$ *such that* $\eta^k \leqq \hat{\eta}$ *for all* k.

Proof. In view of the algorithm's rules, it suffices to show that there exists $\tilde{\eta} > 0$ such that if (4.2) holds, $k \geqq k_E$, and $\eta \geqq \tilde{\eta}$, then $|\bar{d}^k(\eta)|_{\bar{A}} \leqq 1$, where $(\bar{d}^k(\eta), u^k(\eta))$:

$$\text{minimize } \tfrac{1}{2}\eta|\bar{d}|_{\bar{A}}^2 + u \quad \text{over all } (d, u) \in R^{N+1}$$

$$\text{satisfying } f_j(\bar{x}_c) + \langle g^j, \bar{d}\rangle \leqq u \quad \text{for } j \in J^k.$$

Since $\bar{d} = 0$ and $u = f(\bar{x}_c) \geqq f_j(\bar{x}_c)$ for all $j \in J^k$ are feasible above,

$$\tfrac{1}{2}\eta|\bar{d}^k(\eta)|_{\bar{A}}^2 + u^k(\eta) \leqq f(\bar{x}_c). \tag{4.4}$$

The definition of f_j and the Cauchy-Schwarz inequality give

$$\begin{aligned} u^k(\eta) &= \max\{f_j(\bar{x}_c) + \langle g^j, \bar{d}^k(\eta)\rangle : j \in J^k\} \\ &\geqq \min\{f(y^j) + \langle g^j, \bar{x}_c - y^j\rangle : j \in J^k\} - |\bar{d}^k(\eta)| \max_{j \leqq k} |g^j|. \end{aligned}$$

Hence there exist constants $C_1 < f(\bar{x}_c)$ and $C_2 > 0$ such that

$$u^k(\eta) \geqq C_1 - |\bar{d}^k(\eta)|C_2 \quad \text{for all } \eta > 0 \text{ and } k \geqq k_E, \tag{4.5}$$

because $y^j \in E_j = \bar{E}$ if $j > k_E$, $\bar{E}$ is bounded, f is continuous, $g^j \in \partial f(y^j)$, and ∂f is locally bounded. Combining (4.3)-(4.5), we get

$$f(\bar{x}_c) - C_1 \geqq |\bar{d}^k(\eta)|_{\bar{A}}(\tfrac{1}{2}\eta|\bar{d}^k(\eta)|_{\bar{A}} - C_2/\gamma).$$

Therefore, if $|\bar{d}^k(\eta)|_{\bar{A}} > 1$ then $\eta/2 - C_2/\gamma \leqq f(\bar{x}_c) - C_1$, and the existence of $\tilde{\eta}$ is clear.

By the rules of Step 5,

$$x^k = x^{k(L)} \quad \text{if } k(L) \leqq k < k(L+1),$$

where for theoretical purposes we may let $k(L+1) = +\infty$ if the number L of serious steps stays bounded, i.e., if $x^k = x^{k(L)}$ for some fixed L and all $k \geqq k(L)$. First we consider the case of unbounded L.

LEMMA 4.5. *Suppose that Algorithm* 3.1 *executes infinitely many serious steps. Then* $x^k \to \bar{x}_c$ *and* $\bar{x}_c \in X^*$.

Proof. Let $K = \{k(L+1) - 1: L = 1, 2, \cdots\}$, so that at Step 5 $f(x^{k+1}) \leqq f(x^k) + mv^k$ for all $k \in K$. Since $-v^k = |p^k|_{B_k}^2/\eta^k + \bar{\alpha}_p^k > 0$ for $k \geqq k_E$ and $m \in (0, 1)$ is fixed, whereas $f(x^k) \geqq f(x^{k+1}) \geqq \min f$ for all k, passing to the limit with $k \in K$ in the inequality $f(x^k) - f(x^{k+1}) \geqq -mv^k$ yields $\bar{\alpha}_p^k \overset{K}{\to} 0$ and $|p^k|_{B_k}/\eta^k \overset{K}{\to} 0$. Letting $k \in K$ approach infinity in (2.18), we get $f(x^k) \downarrow \min f$. By (4.3), (4.2), and (2.9a), $\gamma|x^{k+1} - \bar{x}_c| \leqq |x^{k+1} - \bar{x}_c|_{\bar{A}} = |y^{k+1} - \bar{x}_c|_{A_k} = |\bar{d}^k|_{A_k} = |p^k|_{B_k}/\eta^k$ for large $k \in K$, so $|p^k|_{B_k}/\eta^k \overset{K}{\to} 0$ implies $x^k \to \bar{x}_c$. Hence $f(x^k) \downarrow f(\bar{x}_c)$ due to the continuity of f, and $f(\bar{x}_c) = \min f$, as desired.

It remains to consider the case of infinitely many successive null steps.

LEMMA 4.6. *Suppose that* $x^k = x^{k(L)} = \bar{x}$ *for some fixed* L *and all* $k \geqq k(L)$. *Then* $\bar{x} \in X^*$ *and* $\bar{x}_c \in X^*$.

Proof. In view of Lemma 4.4, the algorithm's rules imply the existence of $\bar{\eta} > 0$ and $\bar{k} > \max\{k(L), k_E\}$ such that $\eta^k = \bar{\eta}$ for all $k \geqq \bar{k}$; otherwise, successive increases of η^k with $\chi > 1$ at Step 4 would make it unbounded, a contradiction.

Let $k \geqq \bar{k}$ be fixed. Since $y^k = x_c^{k-1} + \bar{d}^{k-1}$, $x^k = x^{k-1} = \bar{x}$, $x_c^k = x_c^{k-1} = \bar{x}_c$ and $f(y^k) > f(x^{k-1}) + mv^{k-1}$, we obtain from (2.6)

$$\bar{\alpha}_k^k = f(\bar{x}) - f(y^k) - \langle g^k, \bar{x}_c - y^k\rangle, \qquad -\bar{\alpha}_k^k + \langle g^k, \bar{d}^{k-1}\rangle > mv^{k-1}. \tag{4.6}$$

Suppose the matrix $\bar{\eta}\bar{A}$ is factorized as $\bar{\eta}\bar{A} = \hat{A}^T\hat{A}$, where $\hat{A}$ is $N \times N$ and nonsingular, and consider the space transformation

$$x \to \hat{x} = \hat{A}x, \tag{4.7}$$

which maps x^k, $\bar{x}$, y^j, and $\bar{d}^k$ into $\hat{x}^k$, $\hat{x}$, $\hat{y}^j$, and $\hat{d}^k$, respectively. Also let $\hat{g}^j = \hat{A}^{-T} g^j$ and $\alpha_j^k = \bar{\alpha}_j^k + \langle g^j, x_c^k - x^k \rangle$ for all $j \in J^k$, $\hat{p}^k = \hat{A}^{-T} p^k$ and $\tilde{\alpha}_p^k = \bar{\alpha}_p^k + \langle p^k, x_c^k - x^k \rangle$.

Thus for $k \geqq \hat{k}$ we have $\eta^k A_k = \hat{A}^T \hat{A}$, $\bar{\alpha}_p^k > 0$ $(k > k_E)$, $\hat{d}^k = -\hat{p}^k$ and $-v^k = |\hat{p}^k|^2 + \bar{\alpha}_p^k$ by (2.9), $w^k = |\hat{p}^k|^2/2 + \bar{\alpha}_p^k$ by (2.10), $\hat{y}^{k+1} = \hat{x}_c^k + \hat{d}^k$ (with $\hat{x}_c^k = \hat{A}\bar{x}_c$), and, by (4.6), $-\bar{\alpha}_k^k + \langle \hat{g}^k, \hat{d}^{k-1} \rangle > m v^{k-1}$ with $0 < m < 1$. Hence we may use the various relations of § 2 to deduce that for all $k \geqq \bar{k}$, Algorithm 3.1 is essentially equivalent to the method of [K2, § 2.5] applied to the convex function $\hat{f}(\cdot) = f(\hat{A}^{-1} \cdot)$ in the space transformed via (4.7). Then the results in [K2, §§ 2.4, 2.5] imply that $w^k = |p^k|_B^2/2\eta^k + \bar{\alpha}_p^k \to 0$ (see (2.10)). Therefore, $|p^k|_{B_k} \to 0$ and $\bar{\alpha}_p^k \to 0$, since $\bar{\alpha}_p^k > 0$ and $\eta^k = \bar{\eta}$ for all $k > \bar{k}$, and (2.18) yields $\bar{x} \in X^*$.

It remains to show that $\bar{x}_c \in X^*$. Since $w^k \to 0$ and $\bar{\alpha}_p^k > 0$ for large k, (2.9b) and (2.10) imply that $v^k \to 0$. Hence $u^k = f(x^k) + v^k \to f(\bar{x})$. On the other hand, by (2.4),

$$u^k \geqq f_k(y^{k+1}) = f(y^k) + \langle g_f(y^k), y^{k+1} - y^k \rangle \tag{4.8}$$

because $k \in J^k$ for all k. Since $|y^{k+1} - \bar{x}_c|_{\bar{A}} = |p^k|_{B_k}/\bar{\eta}$ for large k, and $|p^k|_{B_k} \to 0$, (4.3) implies that $y^k \to \bar{x}_c$. Passing to the limit in (4.8), we get $f(\bar{x}) \geqq f(\bar{x}_c)$, since $u^k \to f(\bar{x})$, $y^k \to \bar{x}_c$, f is continuous and $g_f(\cdot)$ is locally bounded. Thus $\bar{x} \in X^*$ and $f(\bar{x}) \geqq f(\bar{x}_c)$, so $\bar{x}_c \in X^*$, as desired.

We conclude from Lemmas 4.5 and 4.6 that the case of a finite number of ellipsoid updates is rather unlikely, since then the center of the last ellipsoid must be optimal.

Combining Lemmas 4.2–4.6, we deduce our principal result.

THEOREM 4.7. *Either the sequence* $\{x^k\}$ *generated by Algorithm* 3.1 *is finite and its last element minimizes* f, *or* $\{x^k\}$ *is infinite and* $f(x^k) \downarrow \min f$ *as* $k \to \infty$.

The key condition needed to ensure convergence is $E_1 \cap X^* = \emptyset$. Even if it fails, we still have the following theorem.

THEOREM 4.8. *If Algorithm* 2 *is applied to a general convex function* f, *which does not necessarily attain its infimum on* R^N, *then either of the following holds*:

(i) *The sequence* $\{x^k\}$ *is finite and its last element* x^k *minimizes* f *on the set* $\bigcup_{j=1}^k E_j$;

(ii) $\{x^k\}$ *is infinite and* $\lim_{k\to\infty} f(x^k) \leqq \inf\{f(x): x \in \bigcup_{j=1}^\infty E_j\}$ (*e.g.*, $\lim_{k\to\infty} f(x^k) = -\infty$).

Proof. Extend Lemma 4.1 by showing that

$$T(f(x^k)) \cap \bigcup_{j=1}^k E_j \subset E_k \tag{4.9}$$

and use the proof of Lemma 4.3 with $\bar{f} > f(x)$ and an arbitrary $x \in \bigcup E_j$ to deduce that either (i) or (ii) holds if Step 3 is executed infinitely often. In the remaining case, suppose that $\{f(x^k)\}$ is bounded from below and obtain (ii) from the proofs of Lemmas 4.5 and 4.6.

Note that relations (2.16) and (4.9) imply that

$$f(x^k) \leqq \min\left\{f(x): x \in \bigcup_{j=1}^k E_j\right\} + |p^k|_{B_k} + \bar{\alpha}_p^k,$$

which justifies the stopping criterion of Step 2 in the general case.

5. Ellipsoid updating strategies. Proceeding as in [G6], we can establish an upper bound on the rate of convergence in objective values in terms of the rate of volume reduction of successive ellipsoids of the method. To obtain a faster volume reduction, the following modification will use more ellipsoid cuts.

Suppose that at Step 0 we choose a fixed $\bar{\omega} \in (-1/N, 0]$. Steps 2 and 3 are replaced by the following:

Step 2′ (Stopping criterion). If $|p^k|_{B_k} + \bar{\alpha}_p^k \leqq \varepsilon_s$ or $|g^j|_{B_k} + \bar{\alpha}_j^k \leqq \varepsilon_s$ for some $j \in J^k$, terminate; otherwise, continue.

Step 3′ (Ellipsoid updating). (i) Set the ellipsoid update indicator $i_E = 0$.

(ii) Compute the algebraic distances

$$\omega_p^k = -\bar{\alpha}_p^k / |p^k|_{B_k} \quad \text{if } i_E = 0 \quad \text{and} \quad \omega_j^k = o - \bar{\alpha}_j^k / |g^j|_{B_k} \quad \text{for } j \in J^k$$

from x_c^k to H_p^k and H_j^k, respectively (cf. (2.13), (2.17)). Let j^* maximize ω_j^k over $j \in J^k$. If $i_E = 0$ and $\omega_p^k \geqq \omega_{j^*}^k$, set $\omega = \omega_p^k$ and $H = H_p^k$; otherwise, set $\omega = \omega_{j^*}^k$ and $H = H_{j^*}^k$.

(iii) If $\omega < \bar{\omega}$, go to (v); otherwise, continue.

(iv) Let $E_+ = \{x \in R^N : |x - x_c^+|_{A_+} \leqq 1\}$ be the smallest-volume ellipsoid containing $E_k \cap H$. Set $i_E = 1$, replace x_c^k and B_k by x_c^+ and A_+^{-1}, and go to (ii).

(v) If $i_E = 0$, go to Step 4; otherwise, choose $\eta^+ \in (0, \eta_u]$, replace η^k by η^+, and go to Step 1.

The modification above uses "ordinary" cuts based on relations (2.17), which provide the following analogue of (2.18)

$$f(x^k) \leqq \min f + |g^j|_{B_k} + \bar{\alpha}_j^k \tag{5.1}$$

for the stopping criterion. Step 3′(ii) chooses the (possibly nonunique) *deepest* cut if $\omega > 0$, or the *least shallow* cut if $\omega \leqq 0$. The condition $\omega \geqq \bar{\omega} > -1/N$ ensures a significant volume reduction, since at Step 3′(iv) we have $\operatorname{Vol}(E_+) = q(\omega) \operatorname{Vol}(E_k)$ with

$$q(\omega) = \left(\frac{N^2}{N^2-1}\right)^{(N-1)/2} \frac{N}{N+1} (1-\omega^2)^{(N-1)/2}(1-\omega),$$

$$q(\omega) \leqq q(\bar{\omega}) \leqq \bar{q} := e^{-N(\bar{\omega}+1/N)^2/3} < 1$$

for $N > 1$ (see [T1], [G7]). On the other hand, $\omega < \bar{\omega}$ implies that

$$|\bar{\omega}| E_k \subset T^k \subset E_k, \tag{5.2}$$

where $T^k = E_k \cap \bigcap_{j \in J^k} H_j^k$ is an outer approximation to $E_k \cap T(f(x^k))$. If we had $\bar{\omega} = -1/N$ and $T^k = T(f(x^k))$, relation (5.2) would mean that the matrix A_k of E_k is a generalized "Hessian" of [G7]. (This terminology comes from the fact that for a smooth convex f with a minimizer x^* the classical Hessian at x^* is associated with the limit of the smallest ellipsoid containing (or the largest ellipsoid contained in) $[T(f(x^*) + \varepsilon^2/2) - x^*]/\varepsilon$ as $\varepsilon \downarrow 0$; see [G7] for details). Wanting E_k to be close to T^k, and hence to $T(f(x^k))$, we would like to have a large value of $|\bar{\omega}|$ in (5.2). Since such a value cannot, in general, be greater than $1/N$ (see [G7]), whereas $\bar{\omega}$ close to $-1/N$ may result in exceedingly many updates, in practice we use $\bar{\omega} = -1/2N$.

The *deepest aggregate* (or surrogate) cut is defined if x_c^k is cut off by at least one half-space H_j^k with $\bar{\alpha}_j^k > 0$. This cut is given by the half-space $\hat{H}^k$ containing $\bigcap_{j \in J^k} H_j^k$ that is furthest from x_c^k in the metric of $|\cdot|_{A_k}$. $\hat{H}^k$ can be found by projecting x_c^k on $\bigcap_j H_j^k$. Thus we may find for d_H^k:

$$\begin{aligned} &\text{minimize } \tfrac{1}{2}|d|_{A_k}^2, \\ &\text{subject to } -\bar{\alpha}_j^k + \langle g^j, d\rangle \leqq 0 \quad \text{for } j \in J^k, \end{aligned} \tag{5.3}$$

since $\hat{H}^k = \{x: \langle -A_k d_H^k, x - x_c^k \rangle \leqq -|d_H^k|_{A_k}^2\}$. Equivalently, we may find the Lagrange multipliers $\lambda_{H,j}^k$ of (5.3) that

$$\text{minimize } \frac{1}{2}\left|\sum_{j\in J^k} \lambda_j g^j\right|_{B_k}^2 + \sum_{j\in J^k} \bar{\alpha}_j^k,$$

$$\text{subject to } \lambda_j \geqq 0 \quad \text{for } j \in J^k,$$

and let

$$(p_H^k, \alpha_H^k) = \sum_{j\in J^k} \lambda_{H,j}^k (g^j, \bar{\alpha}_j^k),$$

since then $d_H^k = -B_k p_H^k$ and

$$\hat{H}^k = \{x \in R^N: \langle p_H^k, x - x_c^k \rangle \leqq \alpha_H^k\}. \tag{5.4}$$

Of course, finding the deepest aggregate cut involves additional work in quadratic programming. Therefore, it is worthwhile to observe that the usual aggregate cut may be close to the deepest one. For instance, $\hat{H}^k = H_p^k$ if $v^k = 0$. Indeed, in this case subproblems (2.5) and (5.3) produce $\bar{d}^k = d_H^k$, $p^k = p_H^k$, and $\bar{\alpha}_p^k = \alpha_H^k$, so that relations (2.13) and (5.4) imply that $H_p^k = \hat{H}^k$. This suggests that H_p^k approximates $\hat{H}^k$ whenever $|v^k|$ is small relative to $|g^j|_{B_k}$ for $j \in J^k$.

Note that at Step 3′ we could go from (iv) to (v) directly.

There exists an additional possibility for updating the ellipsoid after a serious step. When $x^{k+1} \neq x^k$ at Step 7, we may compute $\bar{\alpha}_j^{k+1}$, $j \in J^k$, $\bar{\alpha}_p^{k+1} = \bar{\alpha}_p^k + f(x^{k+1}) - f(x^k)$, increase k by 1, and, before going to Step 1, execute Steps 2′ and 3′.

So far we have restricted our discussion to single cuts. To obtain a greater reduction in volume, at Step 3′(iv) we may choose a subset J_H^k of J^k such that $\omega_j^k \geqq \bar{\omega}$ for some $j \in J_H^k$, and then let E_+ be the smallest-volume ellipsoid containing the set $S^k = E_k \cap \bigcap_{j\in J_H^k} H_j^k$. In practice the construction of E_+ may be too complicated if J_H^k has more than two elements (see [S3]); for two elements explicit formulae are given in [G1], [E1]. In fact, it is not absolutely necessary to use minimum volume ellipsoids, and we may let E_+ be any ellipsoid containing S^k such that $\text{Vol}(E_+) \leqq q(\bar{\omega})\,\text{Vol}(E_k)$. This extra freedom may facilitate the construction of E_+.

It is straightforward to verify that all the modifications discussed in this section are covered by the convergence analysis of § 4.

We conclude that the algorithm can use a variety of techniques for updating its ellipsoids. Naturally, the best strategy is open to question.

6. Implementation. In this section we discuss our implementation of the algorithm. As in [G8], at the kth iteration we use the factorization

$$B_k = LDL^T$$

with L a lower triangular $N \times N$ matrix with unit diagonal and $D = \text{diag}(d_1, \cdots, d_N)$ a diagonal matrix with positive diagonal. Let $\tilde{B}_k = LD^{1/2}$, so that $B_k = \tilde{B}_k \tilde{B}_k^T$. We do not store the subgradients g^j, but update the vectors

$$\tilde{g}^j = \tilde{B}_k^T g^j = D^{1/2} L^T g^j \quad \text{for } j \in J^k. \tag{6.1}$$

Step 1 is implemented with the quadratic programming routine of [K3]. This routine solves the following version of the dual subproblem (2.11):

$$\begin{aligned} &\text{minimize } \frac{1}{2}\left|\sum_{j\in J^k}\lambda_j\tilde{g}^j\right|^2+\sum_{j\in J^k}\lambda_j\eta^k\bar{\alpha}_j^k,\\ &\text{subject to } \sum_{j\in J^k}\lambda_j=1,\quad \lambda_j\geqq 0\quad \text{for } j\in J^k. \end{aligned} \tag{6.2}$$

It also computes, as byproducts, the quantities

$$(\tilde{p}^k,\bar{\alpha}_p^k)=\sum_{j\in J^k}\lambda_j^k(\tilde{g}^j,\bar{\alpha}_j^k), \tag{6.3}$$

$|p^k|_{B_k}^2=|\tilde{p}^k|^2$, and $\eta^k v^k=-\{|\tilde{p}^k|^2+\eta^k\bar{\alpha}_p^k\}$ (cf. (2.9b)). Hence we can calculate $|\bar{d}^k|_{A_k}^2=|\tilde{p}^k|^2/(\eta^k)^2$ and

$$\bar{d}^k=-(1/\eta^k)\tilde{B}_k\tilde{p}^k=-(1/\eta^k)LD^{1/2}\tilde{p}^k. \tag{6.4}$$

The algorithm terminates at Step 2′ if

$$\min\{|\tilde{p}^k|+\bar{\alpha}_p^k,|\tilde{g}^j|+\bar{\alpha}_j^k: j\in J^k\}\leqq\varepsilon_s(1+|f(x^k)|), \tag{6.5}$$

where $\varepsilon_s>0$ is the desired relative final accuracy in the objective value, i.e., $f(x^k)-\min f\leqq\varepsilon_s(1+|f(x^k)|)$ at termination; see (2.18), (5.1). (This stopping criterion is less sensitive to problem scaling than that of Algorithm 3.1.) Hence we compute the norms $|\tilde{g}^j|$, which are also needed by the quadratic programming routine.

For simplicity, we use the single cuts described at Step 3′ in § 5.

The ellipsoid updating at Step 3′(iv) is trivial if $N=1$. Hence, suppose that $N>1$ and H_p^k is used to cut E_k. As in [G8], we compute $\omega_p^k=-\bar{\alpha}_p^k/|\tilde{p}^k|$,

$$\begin{aligned} &\omega=\min\{\omega_p^k,1-\varepsilon_M\},\\ &t_c=(1+N\omega)/(N+1)|\tilde{p}^k|,\\ &x_c^+=x_c^k-t_cLD^{1/2}\tilde{p}^k,\\ &\delta=N^2(1-\omega^2)/(N^2-1),\\ &\sigma=2(1+N\omega)/(N+1)(1+\omega), \end{aligned} \tag{6.6}$$

where the use of the relative machine precision ε_M ensures that the updated matrix $B_+=A_+^{-1}$ given by

$$B_+=\delta[B_k-\sigma B_kp^k(B_kp^k)^T/|p^k|_{B_k}^2]$$

can be factorized as $B_+=L_+D_+L_+^T$ with L_+ lower triangular and D_+ diagonal with positive diagonal. More specifically, we compute

$$\gamma=D^{1/2}\tilde{p}^k/|\tilde{p}^k|,$$

so that $B_+=L[\delta(D-\sigma\gamma\gamma^T)]L^T$, and then find a diagonal matrix $\tilde{D}=\operatorname{diag}(\tilde{d}_1,\cdots,\tilde{d}_N)$ and a unit lower triangular matrix $\tilde{L}=(\tilde{l}_{ij})$ such that $D-\sigma\gamma\gamma^T=\tilde{L}\tilde{D}\tilde{L}^T$ by using the following recurrence relations:

(i) Set $t_{N+1}=(N-1)/(N+1)(1-\omega)/(1+\omega)$;

(ii) For $j=N-1,\cdots,1$ set

$$\begin{aligned} &t_j=t_{j+1}+\sigma\gamma_j^2/d_j,\\ &\tilde{d}_j=d_jt_{j+1}/t_j,\\ &\beta_j=-\sigma\gamma_j/(d_jt_{j+1}) \end{aligned}$$

and by setting $\tilde{l}_{ij} = \gamma_i \beta_j$ for $j < i$ (see [G3]). Then $D_+ = \delta \tilde{D}$ and the product $L_+ = (l_{ij}^+) V = L\tilde{L}$ can be computed from the recurrence relations (see [G2]):

(i) Set $q_N^N = \gamma_N$;

(ii) For $j = N-1, \cdots, 1$ set

$$\begin{aligned} q_j^j &= \gamma_j, \\ l_{ij}^+ &= l_{ij} + \beta_j q_i^{i+1} \\ q_i^j &= q_i^{j+1} + \gamma_j l_{ij} \end{aligned} \quad \text{for } i = j+1, \cdots, N.$$

To update the quantities $\bar{\alpha}_j^k = f(x^k) - f_j(x_c^k)$ we use the relations

$$f_j(x_c^+) - f_j(x_c^k) = \langle g^j, x_c^+ - x_c^k \rangle = \langle \tilde{g}^j, \tilde{B}_k^{-1}(x_c^+ - x_x^k) \rangle = -t_c \langle \tilde{g}^j, \tilde{p}^k \rangle,$$

which follow from (6.1) and (6.6). Next, to update the transformed subgradients (6.1), we note that

$$D_+^{1/2} L_+^T g^j = D_+^{1/2} \tilde{L}^T L^T g^j = D_+^{1/2} \tilde{L}^T D^{-1/2} \tilde{g}^j,$$

so that $\bar{g}^j = D_+^{1/2} L_+^T g^j$ can be computed from the backward recurrence (see [G2]):

(i) Set $s = 0$, $\bar{g}_N^j = \tilde{d}_N \tilde{g}_N^j$;

(ii) For $i = N, N-1, \cdots, 2$ set

$$\begin{aligned} s &= s + \tilde{\gamma}_i \tilde{g}_i^j, \\ \bar{g}_{i-1}^j &= \tilde{d}_{i-1} \tilde{g}_{i-1}^j + \tilde{\beta}_{i-1} s, \end{aligned}$$

where the quantities $\tilde{d}_i = (d_i^+ / d_i)^{1/2}$, $\tilde{\gamma}_i = \gamma_i / d_i^{1/2}$, $\tilde{\beta}_i = \beta_i (d_i^+)^{1/2}$ are computed beforehand, using the stored values of $d_i^{1/2}$ and $(d_i^+)^{1/2}$.

The convergence analysis of § 4 imposes only weak restrictions on the choice of $\{\eta^k\}$. Relation (5.2) with $|\bar{\omega}| = \frac{1}{2}N$ suggests that it may be useful to restrict the trial point finding to the smaller ellipsoid $|\bar{\omega}| E_k$ which, being an inner approximation to T^k, should be almost contained in $T(f(x^k))$. Hence our strategy for selecting η^k aims for y^{k+1} to

$$\begin{aligned} &\text{minimize } \hat{f}^k(y) \\ &\text{subject to } y \in r_e E_k \end{aligned}$$

or equivalently for $\bar{d}^k = y^{k+1} - x_c^k$ to

$$\begin{aligned} &\text{minimize } \hat{f}^k(x_c^k + \bar{d}) \\ &\text{subject to } |\bar{d}|_{A_k} \leqq r_e, \end{aligned} \tag{6.7}$$

where $r_e \in (0, 1]$ is a trust region radius. We use the fixed value $r_e = 1/2N$ (corresponding to $\bar{\omega} = -1/2N$), which seems to work better than $r_e = 1$, although an adaptive choice of r_e at each iteration could be more efficient.

We relate subproblems (6.2) and (6.7) through the following choice of η^k. At Steps 3'(v) and 7 we set η^+ and η^{k+1} (if $x^{k+1} \neq x^k$) to the value of $\eta_{\min} = 10^{-5}$, since smaller values may lead to inaccuracies at quadratic programming (cf. (6.4)). At Step 4, if $|\bar{d}^k|_{A_k} > 1.2 r_e$, η^k is increased to

$$\eta^+ = \max \{1.2 |\bar{d}^k|_{A_k} / r_e, \chi\} \eta^k, \tag{6.8}$$

where $\chi > 1$ is a parameter. When η^+ replaces η^k in (6.2), the corresponding $\tilde{p}^+$ and $\bar{d}^+$ given by (6.3) and (6.4) satisfy $|\tilde{p}^+| \geqq |\tilde{p}^k|$, $|\bar{d}^+|_{A_k} = |\tilde{p}^+| / \eta^+$ and

$$|\bar{d}^+|_{A_k} \geqq |\bar{d}^k|_{A_k} \eta^k / \eta^+ \geqq \min \{1/1.2, 1.2/\chi\} r_e.$$

Hence, eventually we get

$$\min\{1/1.2, 1.2/\chi\} r_e \leqq |\bar{d}^k|_{A_k} \leqq 1.2 r_e, \tag{6.9}$$

which implies satisfaction of the trust region constraint of (6.7) with the relative accuracy of 20 percent if $\chi = 1.5$. It might appear that, since η^k is not decreased after null steps with no ellipsoid updates, we could have $|\bar{d}^k|_{A_k}$ much smaller than r_e at Step 4. Yet in our calculations the bound (6.9) was (slightly) violated in negligibly few cases.

We could, of course, consider other ways of selecting η^k, e.g., safeguarded interpolation using parametric analysis of subproblem (6.2) with respect to η^k. We note that when the change of η^k is small enough, our quadratic programming routine [K3] can solve the new subproblem very quickly by exploiting the information gathered so far.

We use the following subgradient selection strategy. At Step 0 we choose the maximum number $M_g \geqq N+2$ of stored subgradients. At Step 6 we initially set $J^{k+1} = J^k \cup \{k+1\}$ and then, if necessary, drop from J^{k+1} an index $j \notin \hat{J}^k$ with the largest value of $f(x^{k+1}) - f_j(x_c^k)$, so that J^{k+1} has at most M_g elements.

It is worth adding that, in theory, the algorithm can be made invariant with respect to the objective scaling. To this end, consider the following version. At Step 0 we choose $r_e \in (0, 1]$ and set $\eta^1 = |\tilde{p}^1|/r_e$. At Step 2′ we delete the "1" from the stopping criterion (6.5). At Steps 3′(v) and 7 we set η^+ and η^{k+1} (if $x^{k+1} \neq x^k$) equal to $\eta_u \eta^k$. At Step 4 η^k is not changed if $|\bar{d}^k|_{A_k} \leqq 1.2 r_e$; otherwise, η^+ is calculated from (6.8). Moreover, suppose that the arbitrary decisions of Steps 3′(ii) and 6, concerning the selection of cuts and subgradients, are made according to some fixed rules that account for the possible ties. Note that the proof of Lemma 4.4 can be extended to cover this version.

The described version above is scale invariant in the following sense. Applying the algorithm to f, we get sequences $\{x^k\}$, $\{\bar{d}^k\}$, $\{A_k\}$, $\{\eta^k\}$, etc. Next, suppose we use the same parameters at Step 0 and apply the algorithm to sf, where $s > 0$ is fixed, with the subgradient mapping sg_f to get sequences $\{x_s^k\}$, $\{\bar{d}_s^k\}$, $\{A_k^s\}$, $\{\eta_s^k\}$. Then $x_s^k = x^k$, $d_s^k = \bar{d}^k$, $A_k^s = A_k$, and $\eta_s^k = s\eta^k$ for all k such that termination does not occur before iteration k. To save space, we omit an inductive proof of this fact. (Some hints for the proof: multiply the objective of (6.2) by s^2 and relations (6.3), (6.8), (2.4), and (2.19) by s, observe that the division in (6.4) cancels s, and use the uniqueness of $(\hat{p}^k, \bar{\alpha}_p^k)$ and relations (2.13) and (2.17) for ellipsoid updates.)

7. Subgradient aggregation. The algorithm described so far will typically use $N+2$ past subgradients at each iteration. We now show how the subgradient aggregation strategy of [K2] can be used to trade off storage and work per iteration for speed of convergence.

At Step 0 we choose the maximum number $M_g \geqq 1$ of stored "ordinary" subgradients. The aggregate subgradient is a convex combination of "ordinary" subgradients, which is generated recursively as follows. At Step 0 we define $p^0 = g^1$ and $\tilde{f}^0(x) = f(y^1) + \langle p^0, x - x^1 \rangle$ for all x. At Step 1 we append to subproblem (2.5) the aggregate constraint

$$-\bar{\alpha}_p^{k-1} + \langle p^{k-1}, \bar{d} \rangle \leqq v, \tag{7.1}$$

where $\bar{\alpha}_p^{k-1} = f(x^k) - \tilde{f}_p^{k-1}(x_c^k)$, and use the Lagrange multiplier λ_p^k of (7.1) to define

$$(p^k, \bar{\alpha}_p^k) = \sum_{j \in J^k} \lambda_j^k (g^j, \bar{\alpha}_j^k) + \lambda_p^k (p^{k-1}, \bar{\alpha}_p^{k-1}).$$

As before, we define $\tilde{f}^k(x) = f(x^k) - \bar{\alpha}_p^k + \langle p^k, x - x_c^k \rangle$ to close the recursion. Then at Step 6 we may let J^{k+1} contain $k+1$ as well as any other $M_g - 1$ past indices, e.g., indices j with the largest values of ω_j^k.

Reasoning as in [K2, § 2.4], we may verify that the convergence results of § 4 remain valid for the version with subgradient aggregation.

8. Numerical examples. We shall now report on computational testing of the algorithm using a double precision Fortran code on an IBM PC/XT clone microcomputer with relative accuracy $\varepsilon_m = 2.2 \times 10^{-16}$ ($= 2.2E-16$).

The parameters of the algorithm had the values $m = 0.1$, $\eta_u = 1E-5$, $\chi = 2$, $\bar{\omega} = -1/2N$ (cf. (5.2)), and $r_e = 1/2N$ (cf. (6.7)).

Table 8.1 summarizes results for several standard test problems taken from the literature, which are reviewed below. The following notation is used. N is the number of variables, M_g is the maximum number of stored past subgradients, x^* is the known solution of a problem, x^1 is the standard starting point, $V^1 = \mathrm{Vol}\,(E_1)$, k is the iteration number at termination, $V^k = \mathrm{Vol}\,(E_k)$ is the volume of the final ellipsoid, and m_e is the total number of ellipsoid updates. The stopping criterion (6.5) was used with a value of ε_s chosen for each problem so as to facilitate comparison with results reported in the literature. For each run of the algorithm, an initial ellipsoid was specified by choosing positive tolerances δ_i such that $|x_i^1 - x_i^*| \leqq \delta_i$ and setting $d_i = N\delta_i^2$ for $i = 1, \cdots, N$ in the initial matrix $B_1 = \mathrm{diag}\,(d_1, \cdots, d_N)$. Typically, three runs with increasing δ_i (denoted by δ^a, δ^b, and δ^c) are reported for each problem.

Example 8.1. The Shor problem has $N = 5$,

$$f(x) = \max\left\{ b_i \sum_{j=1}^{5} (a_{ij} - x_j)^2 : i = 1, \cdots, 10 \right\},$$

$$x^* = (1.12434, 0.97945, 1.47770, 0.92023, 1.12429),$$

$x^1 = 0, 0, 0, 0, 1)$, $\delta_i^{1,a} = |x_i^1 - x_i^*|$ for all i, $\delta_i^{1,b} = 2$ for all i, $\delta_i^{1,c} = 10$ for all i (see [S1, p. 137] for the data a_{ij} and b_i).

Example 8.2. The first Colville problem has $N = 5$,

$$f(x) = \sum_{j=1}^{5} d_j x_j^3 + \sum_{i=1}^{5} \sum_{j=1}^{5} c_{ij} x_i x_j + \sum_{j=1}^{5} e_j x_j + 50 \max\{F(x), 0\},$$

$$F(x) = \max\left\{ b_i - \sum_{j=1}^{5} a_{ij} x_j : i = 1, \cdots, 10 \right\},$$

$x^* = (0.3, 0.3335, 0.4, 0.4285, 0.224)$, $x^1 = (0, 0, 0, 0, 1)$, $\delta_i^{2,a} = |x_i^1 - x_i^*|$, $\delta_i^{2,b} = 1$ and $\delta_i^{2,c} = 10$ for all i (see, e.g., [K2, p. 350] for the problem data).

Example 8.3. The Rosen-Suzuki problem has $N = 4$,

$$f(x) = x_1^2 + x_2^2 + 2x_3^2 + x_4^2 - 5x_1 - 5x_2 - 21x_3 + 7x_4 + F(x),$$

$$F(x) = 5 \max\{F_1(x), F_2(x), F_3(x), 0\},$$

$$F_1(x) = x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_1 - x_2 + x_5 - x_4 - 8,$$

$$F_2(x) = x_1^2 + 2x_2^2 + x_3^2 + 2x_4^2 - x_1 - x_4 - 10,$$

$$F_3(x) = x_1^2 + x_2^2 + x_3^2 + 2x_1 - x_2 - x_4 - 5,$$

$x^* = (0, 1, 2, -1)$, $x_1 = (0, 0, 0, 0)$, $\delta^{3,a} = (1E-4, 1, 2, 1)$, $\delta_i^{3,b} = 2$ and $\delta_i^{3,c} = 10$ for all i.

TABLE 8.1
Test results of the descent ellipsoid method.

Problem	N	M_g	$f(x^*)$	ε_s	$f(x^1)$	δ	$\lvert x^1-x^*\rvert_{A_1}$	V^1	k	$f(x^k)$	V^k	m_e
Shor	5	10	22.60016	$1E-6$	80	$\delta^{1,a}$	1	54.8	45	22.60017	$2E-20$	728
						$\delta^{1,b}$	0.51	9416	47	22.60017	$9E-19$	788
						$\delta^{1,c}$	0.10	$2.9E+7$	52	22.60017	$2E-19$	872
Colville	5	10	−32.3487	$1E-5$	20	$\delta^{2,a}$	1	3.92	46	−32.3485	$9E-20$	650
						$\delta^{2,b}$	0.48	294	45	−32.3486	$6E-21$	747
						$\delta^{2,c}$	0.05	$2.9E+7$	54	−32.3486	$1E-21$	961
Rosen-Suzuki	4	10	−44	$1E-5$	0	$\delta^{3,a}$	0.87	0.016	20	−43.9998	$8E-11$	182
						$\delta^{3,b}$	0.61	1263	29	−44.0000	$5E-11$	355
						$\delta^{3,c}$	0.12	$7.9E+5$	33	−43.9999	$7E-11$	420
MAXQUAD	10	20	−0.8414	$1E-4$	0	$\delta^{4,a}$	1	$1.3E-7$	58	−0.84135	$1E-20$	1122
						$\delta^{4,b}$	0.36	2.55	75	−0.84132	$5E-20$	1749
						$\delta^{4,c}$	0.06	$2.6E+8$	102	−0.84132	$6E-20$	2440
MXHILB	30	45	0	$1E-6$	3.995	$\delta^{5,a}$	0.2	$2.9E+38$	15	$1.3E-8$	$5E+28$	2497
	50	60	0	$1E-6$	4.499	$\delta^{5,b}$	0.2	$4.6E+64$	19	$1.3E-7$	$4E+38$	3114
L1HILB	30	45	0	$1E-6$	41.09	$\delta^{6,a}$	0.2	$2.9E+38$	16	$7.7E-9$	$5E+16$	4698
	50	60	0	$1E-6$	68.82	$\delta^{6,b}$	0.2	$4.6E+64$	23	$3.6E-7$	$7E+33$	3579

Example 8.4. Lemaréchal's MAXQUAD problem has $N = 10$,

$$f(x) = \max \{\langle A^L x, x\rangle - \langle b^L, x\rangle : L = 1, \cdots, 5\},$$
$$A_{ij}^L = A_{ji}^L = \exp(i/j)\cos(i \cdot j)\sin(L), \quad i \neq j,$$
$$A_{ii}^L = i|\sin(i)|/10 + \sum_{j \neq i} |A_{ij}^L|,$$
$$b_i^L = \exp(i/L)\sin(i \cdot L).$$

(This problem is difficult to quote: misprints abound in [K2, pp. 346-347], [L2, p. 151], [Z1, Ex. 7.2], etc.). In [L2, p. 152] the optimum value $f(\bar{x}) = -0.8414$ is quoted for a point $\bar{x}$ which in fact has $f(\bar{x}) = -0.8411$. The best point known to us is

$$x^* = (-0.126257, -3.43783E-2, -6.85716E-3, 2.63606E-2, 6.72949E-2,$$
$$-0.278400, 7.42187E-2, 0.138524, 8.4031E-2, 3.85804E-2)$$

with $f(x^*) = -0.841408$. We used $x^1 = 0$, $\delta_i^{4,a} = |x_i^*|$, $\delta_i^{4,b} = 0.3162$ (corresponding to $d_i = 1$) and $\delta_i^{4,c} = 2$ for all i.

Example 8.5. This academic problem MXHILB with

$$f(x) = \max\left\{\left|\sum_{j=1}^{N} x_i/(i+j-1)\right| : i = 1, \cdots, N\right\}, \qquad x^* = 0,$$

corresponds to solving the equation $Ax = b$, where $b = 0$ and A is an $N \times N$ section of the Hilbert matrix. We used $x^1 = (1, \cdots, 1)$ and $\delta_i^{5,a} = 5$ for all i with $N = 30$ and $N = 50$.

Example 8.6. Problem L1HILB with

$$f(x) = \sum_{j=1}^{N} \left|\sum_{i=1}^{N} x_i \Big/ (i+j-1)\right|, \qquad x^* = 0,$$

is a more difficult version of MXHILB, since it has 2^N linear pieces, whereas MXHILB has $2N$. We used $x_i^1 = 1$, $\delta_i^{6,a} = 5$ for $i = 1, \cdots, N = 30, 50$.

Example 8.7. Lemaréchal's SHELL DUAL problem [L2, p. 154] has a highly nonconvex objective function f of 15 variables, with $f(x^*) = 32.3488$ for

$$x^* = (0.3, 0.3335, 0.4, 0.4283, 0.224, 0, 0, 5.1741, 0, 3.0611, 11.8396, 0, 0, 0.1039, 0).$$

This problem seems to be very difficult for general-purpose descent methods (see [L1]). To tackle nonconvexity, we have incorporated in the algorithm the two-point line search of [K2, p. 103] (with parameters $\bar{t} = 0.01$, $\xi = 0.1$, $\gamma = 10$). To this end, we computed the search direction $d^k = \bar{d}^k + x_c^k - x^k$ from x^k, and the line search procedure found two stepsizes t_L^k and t_R^k, $0 \leqq t_L^k \leqq t_R^k \leqq 1$, the next iterate $x^{k+1} = x^k + t_L^k d^k$, and the next trial point $y^{k+1} = x^k + t_R^k d^k$ that provided the next linearization f_{k+1} of f. Of course, the algorithm is not guaranteed to minimize a nonconvex f. Nevertheless, for $x_i^1 = 1E-4$, $i \neq 12$, $x_{12}^1 = 60$ and

$$\delta = (0.3, 0.3335, 0.4, 0.4283, 0.224, 0.1, 0.1, 10, 0.1, 5, 15, 100, 0.1, 1, 0.1)$$

($f(x^1) = 2400$, $|x^1 - x^*|_{A_1} = 0.66$, $V^1 = 7.3E+5$), the algorithm did converge to an approximate solution, although the usual stopping criterion (6.5) did not work with $\varepsilon_s = 1E-4$ and termination occurred due to exceeding the function evaluation limit of 600. Table 8.2 illustrates the algorithm's progress, with NFEV denoting the total number of function and subgradient evaluations.

TABLE 8.2
SHELL DUAL *problem.*

k	$f(x^k)$	NFEV	V^k	m_e
70	32.6369	153	$2.0E-2$	582
78	32.5482	178	$5.7E-3$	606
97	32.4408	237	$2.3E-4$	654
107	32.4045	268	$9.8E-5$	669
110	32.3929	277	$7.4E-5$	677
112	32.3843	283	$6.4E-5$	679
118	32.3748	306	$3.7E-5$	690
127	32.3648	345	$8.1E-6$	709
165	32.3548	496	$5.9E-9$	793
191	32.3538	600	$8.0E-11$	845

Our results for Examples 8.1-8.4 and 8.7 may be compared with those in [S1, p. 139], [L1], [Z1], and [K2, pp. 346-349]. Additionally, Table 8.3 gives results for f scaled by $s = 0.01$, 1, and 100. They were produced by the usual bundle method of [K2, Chap. 2], the present method and the shifted ellipsoid method suggested by a referee. This third method is obtained from Algorithm 3.1 by setting $x_c^1 = x^1$ at Step 0,

TABLE 8.3
Comparative results for scaled problems.

	Scaling	Usual bundle method		Ellipsoid bundle method		Shifted ellipsoid method	
Problem	s	k	$f(x^k)$	k	$f(x^k)$	k	$f(x^k)$
Shor	1	53	22.60016	45	22.60017	51	22.60018
	100	90	22.60016	44	22.60017	57	22.60016
	0.01	71	22.60016	41	22.60017	52	22.60017
Colville	1	60	−32.3487	46	−32.3485	43	−32.3484
	100	54	−32.3487	42	−32.3484	44	−32.3486
	0.01	29	−32.3486	54	−32.3487	44	−32.3486
Rosen-Susuki	1	44	−44.0000	20	−43.9998	30	−44.0000
	100	48	−43.9998	20	−43.9997	30	−44.0000
	0.01	76	−43.9999	19	−43.9994	32	−44.0000
MAXQUAD	1	95	−0.84133	58	−0.84135	94	−0.84140
	100	318	−0.84138	65	−0.84138	98	−0.84140
	0.01	59	−0.84138	78	−0.84140	96	−0.84140
MXHILB	1	590	$5.1E-6$	15	$1.3E-8$	65	$4.5E-10$
$N = 30$	100	2000[1]	$1.8E-5$	18	$5.4E-7$	72	$1.1E-7$
	0.01	1000[1]	$4.0E-3$	15	$2.1E-7$	65	$3.8E-7$
MXHILB	1	500[1]	$1.6E-4$	19	$1.2E-7$	220	$2.4E-9$
$N = 50$	100	46	$1.2E-7$	19	$2.5E-10$	221	$7.9E-9$
	0.01	5000[1]	$2.8E-3$	19	$2.2E-7$	223	$1.8E-7$
L1HILB	1	132	$1.6E-6$	16	$7.7E-9$	66	$1.1E-8$
$N = 30$	100	38	$1.2E-7$	16	$5.1E-7$	151	$1.5E-7$
	0.01	500[1]	$2.3E-2$	15	$3.8E-7$	67	$9.8E-8$
L1HILB	1	199	$8.4E-7$	23	$3.6E-7$	184	$1.1E-9$
$N = 50$	100	46	$1.2E-7$	25	$2.9E-9$	276	$5.1E-7$
	0.01	500[1]	$5.8E-4$	26	$7.8E-8$	187	$2.9E-8$

[1] Termination due to the iterations limit.

deleting Step 3 and setting $x_c^{k+1} = x^{k+1}$ and $A_{k+1} = A_k$ at Step 7; in effect, the initial ellipsoid trust region only shifts with its center x^k, but does not shrink. Using the same line search procedure for handling nonconvexity, the three methods were applied to scaled versions of the Shell Dual; see Tables 8.4-8.6. (For Examples 8.1-8.4 we quote results only for the smallest δ^a. With $\delta = \delta^b$ or δ^c both our method and the shifted ellipsoid method converge more slowly; cf. Table 8.1.)

A comment on the variation in η^k is in order. Usually the final η^k (at Step 5) decreases "smoothly" as k grows, e.g., from 1300 to $6E-4$, 1000 to 0.02, 660 to 0.013, 80 to 0.015, 1200 to $1.5E-5$, 4500 to $1.3E-5$, 8200 to $1E-5$, and from 3900 to $1E-5$

TABLE 8.4
Shell Dual—the usual bundle method.

	Scaling					
	$s = 1$		$s = 100$		$s = 0.01$	
NFEV	k	$f(x^k)$	k	$f(x^k)$	k	$f(x^k)$
100	34	87.97	26	1028	43	2084
200	64	38.84	56	378	85	173
300	94	34.51	84	118	121	158
400	125	33.03	114	43.95	155	145
500	157	32.90	148	35.92	191	133
600	191	32.72	186	33.50	219	128

TABLE 8.5
Shell Dual—the ellipsoid bundle method.

	Scaling					
	$s = 1$		$s = 100$		$s = 0.01$	
NFEV	k	$f(x^k)$	k	$f(x^k)$	k	$f(x^k)$
100	52	33.146	47	32.942	44	34.527
200	85	32.496	78	32.490	77	32.845
300	116	32.378	104	32.410	110	32.563
400	141	32.358	128	32.392	141	32.431
500	166	32.355	152	32.383	171	32.392
600	191	32.354	177	32.371	200	32.372

TABLE 8.6
Shell Dual—the shifted ellipsoid method.

	Scaling					
	$s = 1$		$s = 100$		$s = 0.01$	
NFEV	k	$f(x^k)$	k	$f(x^k)$	k	$f(x^k)$
100	47	37.888	54	33.396	33	812
200	78	34.507	93	33.132	59	493
300	112	33.284	130	32.871	88	308
400	140	33.033	170	32.835	117	149
500	173	32.870	211	32.591	147	50.2
600	198	32.844	243	32.583	177	37.4

for the problems with $s=1$ reported in Table 8.3 for our method, respectively (and from $5.3E+5$ to 70 for the Shell Dual). A similar variation in η^k is observed for the shifted ellipsoid method. Naturally, η^k increases (decreases) for larger (smaller) s.

Of course, no firm conclusions should be drawn from such limited computational experience, but the test results indicate that our method is promising.

9. Conclusions. We have shown how to incorporate an ellipsoid variable metric in a bundle method for convex minimization. The method seems to be promising. We hope to increase its efficiency by using more refined ellipsoid updates.

Acknowledgment. I would like to thank Claude Lemaréchal and two anonymous referees for several helpful suggestions.

REFERENCES

[A1] M. AKGÜL, *Topics in Relaxation and Ellipsoidal Methods*, Research Notes in Mathematics 97, Pitman, Boston, 1984.

[B1] R. B. BLAND, D. GOLDFARB, AND M. J. TODD, *The ellipsoid method: a survey*, Oper. Res., 29 (1981), pp. 1039-1091.

[D1] V. F. DEMYANOV AND L. V. VASILEV, *Nondifferentiable Optimization*, Nauka, Moscow, 1981. (English translation, Optimization Software, Springer-Verlag, Berlin, New York, 1985.)

[E1] A. ECH-CHERIF AND J. G. ECKER, *A class of rank-two algorithms for convex programming*, Math. Programming, 29 (1984), pp. 187-202.

[E2] J. G. ECKER AND M. KUPFERSCHMID, *An ellipsoid algorithm for nonlinear programming*, Math. Programming, 27 (1983), pp. 1-15.

[G1] V. I. GERSHOVICH, *On a cut-off method using linear space transformations*, in Theory of Optimal Solution, Institute of Cybernetics, Kiev, 1979, pp. 15-23. (In Russian.)

[G2] P. E. GILL, G. M. GOLUB, W. MURRAY, AND M. A. SAUNDERS, *Methods for modifying matrix factorizations*, Math. Comput., 28 (1974), pp. 505-535.

[G3] P. E. GILL, W. MURRAY, AND M. A. SAUNDERS, *Methods for computing and modifying the* LDV *factors of a matrix*, Math. Comput., 29 (1975), pp. 1051-1077.

[G4] J. L. GOFFIN, *Variable metric relaxation methods, Part* I: *A conceptual algorithm*, Tech. Report SOL 81-16, Systems Optimization Laboratory, Stanford University, Stanford, CA, 1981.

[G5] ———, *Convergence results in a class of variable metric subgradient methods*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 283-326.

[G6] ———, *Convergence rates of the ellipsoid method on general convex functions*, Math. Oper. Res., 8 (1983), pp. 135-150.

[G7] ———, *Variable metric relaxation methods, Part* II: *the ellipsoid method*, Math. Programming, 30 (1984), pp. 147-162.

[G8] D. GOLDFARB AND M. J. TODD, *Modifications and implementation of the ellipsoid algorithm for linear programming*, Math. Programming, 23 (1982), pp. 1-19,

[H1] J. HALD AND K. MADSEN, *Combined* LP *and quasi-Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49-62.

[K1] J. E. KELLEY, *The cutting plane method for solving convex programs*, J. Soc. Appl. Math., 8 (1960), pp. 703-712.

[K2] K. C. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics 1133, Springer-Verlag, Berlin, New York, 1985.

[K3] ———, *A method for solving certain quadratic programming problems arising in nonsmooth optimization*, IMA J. Numer. Anal., 6 (1986), pp. 137-152.

[L1] C. LEMARÉCHAL, *Numerical experiments in nonsmooth optimization*, in Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., Report CP-82-S8, International Institute for Applied Systems Analysis, Laxenberg, Austria, 1981, pp. 61-84.

[L2] C. LEMARÉCHAL AND R. MIFFLIN, eds., *Nonsmooth Optimization*, Pergamon Press, Oxford, 1978.

[L3] C. LEMARÉCHAL AND J. J. STRODIOT, *Bundle methods, cutting plane algorithms and σ-Newton directions*, in Nondifferentiable Optimization: Motivations and Applications, V. F. Demyanov and D. Pallaschke, eds., Lecture Notes in Economics and Mathematical Systems 255, Springer-Verlag, Berlin, New York, 1985, pp. 25-33.

[L4] C. LEMARÉCHAL AND J. ZOWE, *Some remarks on the construction of higher order algorithms for convex optimization*, J. Appl. Math. Optim., 10 (1983), pp. 51–68.

[M1] R. MIFFLIN, *Better than linear convergence and safeguarding in nonsmooth minimization*, in System Modelling and Optimization, P. Thoft-Christensen, ed., Lecture Notes in Control and Information Sciences 59, Springer-Verlag, Berlin, 1984, pp. 321–230.

[M2] J. J. MORÉ, *Recent developments in algorithms and software for trust region methods*, in Mathematical Programming, The State of the Art, Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 258–287.

[S1] N. Z. SHOR, *Minimization Methods for Non-Differentiable Functions*, Springer-Verlag, Berlin, 1985.

[S2] N. Z. SHOR AND V. I. GERSHOVICH, *Family of algorithms for solving convex programming problems*, Kibernetika, 4 (1979), pp. 62–67. (In Russian.) Cybernetics, 15 (1980), pp. 502–508. (In English.)

[S3] G. SONNEVEND, *A modified ellipsoid method for the minimization of convex functions with superlinear convergence (or finite termination) for well-conditioned C^3 smooth (or piecewise linear) functions*, in Nondifferentiable Optimization: Motivations and Applications, V. F. Demyanov and D. Pallaschke, eds., Lecture Notes in Economics and Mathematical Systems 255, Springer-Verlag, Berlin, New York, 1985, pp. 264–277.

[T1] M. J. TODD, *On minimum volume ellipsoids containing part of a given ellipsoid*, Math. Oper. Res., 7 (1982), pp. 253–261.

[Y1] D. B. YUDIN AND A. S. NEMIROVSKII, *Informational complexity and effective methods for the solution of convex extremal problems*, Ekonom. Mat. Metody, 12 (1976), pp. 357–369. (In Russian.) MATEKON, 13 (1977), pp. 3–25. (In English.)

[Y2] D. B. YUDIN, A. P. GORIASHKO, AND A. S. NEMIROVSKII, *Mathematical Optimization Methods for Devices and Algorithms of Automatic Control Systems*, Radio i Sviaz, Moscow, 1982. (In Russian.)

[Z1] J. ZOWE, *Nondifferentiable optimization*, in Computational Mathematical Programming, K. Schittkowski, ed., Springer-Verlag, Berlin, New York, 1985, pp. 323–356.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 758-775, July 1989

ASYMPTOTIC BEHAVIOR OF SOLUTIONS OF THE ONE-DIMENSIONAL WAVE EQUATION WITH A NONLINEAR BOUNDARY STABILIZER*

HAN-KUN WANG† AND GOONG CHEN‡

Abstract. The modeling of nonlinear passive damping devices or boundary frictions of an otherwise linear vibrating system often results in nonlinear elastic dissipative boundary conditions. Such systems occur increasingly often in engineering applications, whose control and stability analysis appear much more complex than the classical linear distributed parameter systems.

This paper uses the method of characteristics and nonlinear semigroup theory to study the effect of nonlinear boundary stabilization and analyze the asymptotic behavior of solutions of such systems. The authors are able to determine the ω-limit set of the dynamical system and the asymptotic rates of various solutions to the ω-limit set.

Key words. wave equation, nonlinear dissipative boundary condition, ω-limit set

AMS(MOS) subject classifications. 93D15, 93D20, 73D35

1. Introduction. The study of vibration control and suppression has always been an important research area in mechanical and aerospace engineering. Traditionally, as far as passive damping devices are concerned, the most commonly seen ones are probably the viscous dashpots, whose action is to cause a frictional force opposite to the direction of velocity. This frictional force versus velocity relationship is approximately linear, at least within a certain designed operating range. This makes the analysis, design, and control of such linear mechanical systems simple and easily understandable. Nevertheless, such classical linear dashpots are bulky, expensive, and inconvenient to replace and repair.

The advance of modern material science and technology has provided us with useful alternatives such as elastomeric and other viscoelastic damping materials which are generally light weight, durable, and convenient to service. Their utilization in high-performance helicopters, combat aircraft, and vessels in aerospace and naval engineering have sharply increased in the past decade. However, such visco-thermo-elastic materials have highly nonlinear characteristics that cause significant nonlinear response in the entire system. The questions of analysis, design, and control appear more difficult. To the best of our knowledge, research work on such nonlinear energy dissipating devices or materials in a distributed parameter system is rather incomplete.

In this paper, we wish to undertake some research in the above direction. As a modest model, we study a distributed parameter system whose governing equation is linear, viz., a simple one-dimensional wave equation, but the boundary condition is nonlinear dissipative as exemplified by the use of the aforementioned elastomeric material. We will study the effects of vibration and questions involving nonlinear friction. Let us describe our problem below.

Let a vibrating cable or string with mass density m and tension T satisfy the one-dimensional wave equation

$$m\frac{\partial^2 y(x,t)}{\partial t} - T\frac{\partial^2 y(x,t)}{\partial x^2} = 0, \qquad 0<x<1, \qquad t>0. \tag{1.1}$$

* Received by the editors July 15, 1987; accepted for publication (in revised form) November 20, 1988. This research was supported in part by Air Force Office of Scientific Research grants 85-0253 and 87-0334. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation thereon.

† Department of Mathematics, The Wichita State University, Wichita, Kansas 67208.

‡ Department of Mathematics, Texas A&M University, College Station, Texas 77843.

In the equation, $y(x, t)$ denotes the vertical displacement at x at time t, and the length has been normalized to 1. Assume that the string is fixed at the right end:

$$y(1, t) = 0 \quad \text{for all } t > 0. \tag{1.2}$$

At the left end, either some frictional force is present, or a damping device such as a dashpot is installed, as illustrated in Fig. 1. The function of the dashpot is to cause friction, thereby suppressing undesirable vibrations. For an idealized dashpot, the force is assumed to be negatively proportional to velocity at $x = 0$:

$$T\frac{\partial y(0, t)}{\partial x} = k\frac{\partial y(0, t)}{\partial t}, \qquad k > 0, \tag{1.3}$$

where it is known that the left-hand side represents the (negative of) vertical force component at point $x = 0$. If we choose the elastic constant k to be T/c, where $c \equiv (T/m)^{1/2}$ is the wave velocity, then (1.3) becomes

$$\frac{\partial y(0, t)}{\partial t} - c\frac{\partial y(0, t)}{\partial x} = 0, \tag{1.4}$$

which is the characteristic impedance boundary condition. It is known to cause maximum energy loss to the string; the vibration is completely suppressed at time $t = 2/c$.

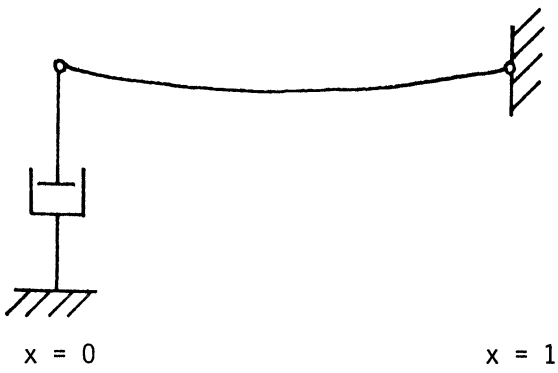

FIG. 1. *A vibrating string with a damping device at left end.*

Now, assume instead that a nonlinear damping device is used whose velocity-frictional force relationship as determined by material testing is as shown in Fig. 2, where a frictional force of magnitude F_0, called the Coulomb friction, is often found present in the nonlinear dashpot (e.g., due to the roughness of the dashpot wall, or due to the aforementioned elastomeric material). The left end point $x = 0$ can be set in motion only after this threshold of frictional force F_0 is overcome by the force exerted by the string at $x = 0$. The velocity range under testing is $[-v_0, v_0]$ for some $v_0 > 0$. Past v_0, plasticity sets in and the string does not behave elastically any more.

We must remark that the response curve as indicated in Fig. 2 has not incorporated the possible effects of *hysteresis*, i.e., the velocity may trace a different curve when it reverses its direction. We wish to address this in a future paper as we know that the hysteresis is an important, commonly observable phenomenon which needs to be carefully studied.

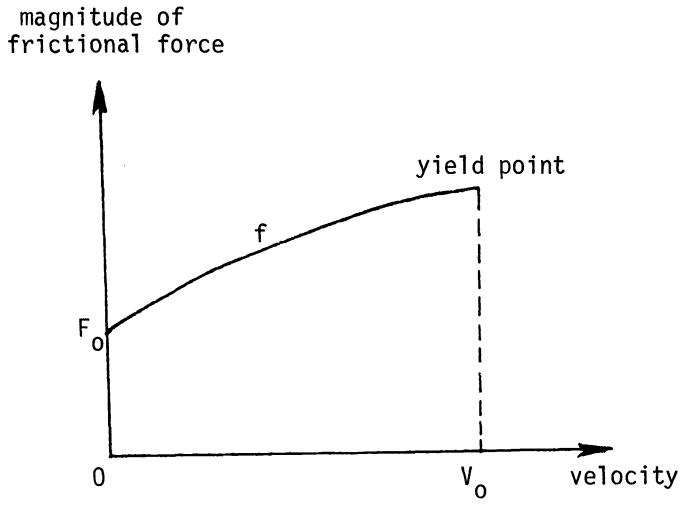

FIG. 2. *Velocity-force relation in a nonlinear dashpot.*

Therefore, we propose the following nonlinear elastic boundary condition as our model.

$$T\frac{\partial y(0,t)}{\partial x} = \bar{f}\left[\frac{\partial y(0,t)}{\partial t}\right], \tag{1.5}$$

where $\bar{f}(x)$ is a multivalued function defined by

$$\bar{f}(\eta) = \begin{cases} -F_0 + g(\eta), & -v_0 < \eta < 0, \\ [-F_0, F_0], & \eta = 0, \\ F_0 + g(\eta), & 0 < \eta < v_0, \end{cases} \tag{1.6}$$

where $[-F_0, F_0]$ denotes the closed interval from $-F_0$ to F_0, and $g(\eta) = f(\eta) - F_0$ for $\eta > 0$, $f(\eta)$ is the function as shown in Fig. 2, $g(\eta)$ satisfies

$$\begin{cases} g(\eta) \text{ is continuous and nondecreasing on } [-v_0, v_0], \\ g(\eta) > 0 \quad \text{if } \eta > 0, \\ g(0) = 0, \\ g(-\eta) = -g(\eta) \quad \text{for } \eta \geqq 0. \end{cases} \tag{1.7}$$

The restriction that the domain of definition of $\bar{f}$ is $[-v_0, v_0]$ is inconvenient. Let us enlarge it by extending g to be an arbitrary function outside $[-v_0, v_0]$ that still satisfies (1.7). Thus,

$$\bar{f}(\eta) = \begin{cases} -F_0 + g(\eta), & \eta < 0, \\ [-F_0, F_0], & \eta = 0, \\ F_0 + g(\eta), & \eta > 0. \end{cases} \tag{1.8}$$

Different extensions of g will not affect the solution provided that the initial data satisfies certain bounds (see the appendix). We also note that the function g may be *nondifferentiable.* Indeed, except for § 4, we do not need the differentiability of g in our treatment.

We wish to study the asymptotic behavior of the solution of (1.1), (1.2), and (1.5), with certain given initial conditions. Our main tools will be the method of characteristics and the nonlinear semigroup theory.

In § 2, we first transform the equation into a hyperbolic system. We show that there is a nonlinear semigroup corresponding to the evolution of the system.

In § 3, we determine the ω-limit set of the dynamical system. A solution may either enter the ω-limit set within finite time, or may circle around the ω-limit set indefinitely without entering it.

The asymptotic rates of convergence to the ω-limit set are determined for several cases in § 4.

2. A nonlinear semigroup of evolution. Define

$$\text{(2.1)}\qquad \begin{cases} \alpha(x,t)=\dfrac{1}{2}\left[\dfrac{\partial y(x,t)}{\partial t}-c\dfrac{\partial y(x,t)}{\partial x}\right] \\ \beta(x,t)=\dfrac{1}{2}\left[\dfrac{\partial y(x,t)}{\partial t}+c\dfrac{\partial y(x,t)}{\partial x}\right]. \end{cases}$$

Then the wave equation (1.1) implies

$$\text{(2.2)}\qquad \begin{cases} \dfrac{\partial}{\partial t}\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix}+\begin{bmatrix} c\dfrac{\partial}{\partial x} & 0 \\ 0 & -c\dfrac{\partial}{\partial x}\end{bmatrix}\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix}=0, \qquad t>0. \\ \text{with initial condition } (\alpha_0(\cdot),\beta_0(\cdot)) \text{ obtained by using} \\ \text{transformation (2.1).} \end{cases}$$

The right boundary condition (1.2) becomes

$$\text{(2.3)}\qquad \beta(1,t)=-\alpha(1,t), \qquad t>0.$$

Consider the left boundary condition (1.5). Since in (1.6), g satisfies $g(0)=0$, for the ease of treatment later on, we define $h(\eta)$ by

$$\text{(2.4)}\qquad h(\eta)=g(\eta)/\eta, \qquad \eta\in\mathbb{R}, \qquad \eta\neq 0.$$

Note that h may be discontinuous at $\eta=0$. From (1.6), using the newly defined h, we derive the following implicit boundary condition:

$$\text{(2.5)}\qquad \alpha(0,t)=\begin{cases} \dfrac{cF_0+[T-ch((\alpha+\beta)(0,t))]\beta(0,t)}{T+ch((\alpha+\beta)(0,t))}, & \beta(0,t)<-\dfrac{cF_0}{2T} \\ -\beta(0,t), & -\dfrac{cF_0}{2T}\leqq\beta(0,t)\leqq\dfrac{cF_0}{2T} \\ \dfrac{-cF_0+[T-ch((\alpha+\beta)(0,t))]\beta(0,t)}{T+ch((\alpha+\beta)(0,t))}, & \beta(0,t)>\dfrac{cF_0}{2T}. \end{cases}$$

For the evolution equation (2.2), the underlying Hilbert space is

$$\mathcal{H}=\left\{\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\,\middle|\,\alpha,\beta\in\mathcal{L}^2(0,1)\right\}=\mathcal{L}^2(0,1)\times\mathcal{L}^2(0,1),$$

whose $\mathcal{L}^2\times\mathcal{L}^2$ product norm is equivalent to the energy of the wave equation. From now on, we will write $\begin{bmatrix}\alpha\\ \beta\end{bmatrix}$ and (α,β) interchangeably, depending on which is convenient.

Let A denote the operator

$$A=\begin{bmatrix} c\dfrac{\partial}{\partial x} & 0 \\ 0 & -c\dfrac{\partial}{\partial x} \end{bmatrix} \tag{2.6}$$

with domain

$$D(A)=\left\{\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\in H^1(0,1)\times H^1(0,1)\,\middle|\,\begin{array}{l}\alpha,\beta \text{ satisfy (2.3) and (2.5)}\\ \text{(with } t \text{ dropped) at } x=0 \text{ and } x=1.\end{array}\right\} \tag{2.7}$$

In (2.7), $H^1(0,1)$ is the standard Sobolev space of order one.

To show that a solution of the problem under consideration exists, we prove that there exists a nonlinear contraction semigroup $S(t)$ corresponding to the dissipative set $-A$, whose minimal section $-A^0$ is the infinitesimal generator. It is known that if $(\alpha_0,\beta_0)\in D(-A^0)$, and if we regard

$$\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix}\equiv S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}$$

as our solution obtained from the semigroup, then $(\alpha(\cdot,t),\beta(\cdot,t))$ has the following smoothness properties:

(i) $(\alpha(\cdot,t),\beta(\cdot,t))\in D(A^0)$ for all $t\geqq 0$, and the function $t\to -A^0(\alpha(\cdot,t),\beta(\cdot,t))$ is continuous from the right on $[0,\infty)$.

(ii) $(\alpha,(\cdot,t),\beta(\cdot,t))$ has a right derivative with respect to t at every $t\geqq 0$ and

$$\frac{d^+}{dt}\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix}=-A^0\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix},\quad \text{for all } t\geqq 0.$$

(iii) $d/dt(\alpha(\cdot,t),\beta(\cdot,t))=-A^0(\alpha(\cdot,t),\beta(\cdot,t))$ exists and is continuous except at a countable number of points $t\geqq 0$.

We refer the reader to [1], [2] for the above and other relevant properties of nonlinear semigroups.

By [1], [2], $-A^0$ generates a nonlinear contraction semigroup if and only if A is maximal monotone. We want to verify this below. As much of the work is rather routine, we will be concise.

LEMMA 1. *A is monotone if and only if g is nondecreasing.*

Proof. For any given (α_1,β_1), $(\alpha_2,\beta_2)\in D(A)$, consider

$$I\equiv\left\langle A\begin{bmatrix}\alpha_1\\ \beta_1\end{bmatrix}-A\begin{bmatrix}\alpha_2\\ \beta_2\end{bmatrix},\begin{bmatrix}\alpha_1-\alpha_2\\ \beta_1-\beta_2\end{bmatrix}\right\rangle,$$

where $\langle\,,\rangle$ is the inner product in $\mathcal{H}$. Using (2.2) and integrating by parts, we get

$$I=\frac{c}{2}[(\beta_1(0)-\beta_2(0))^2-(\alpha_1(0)-\alpha_2(0))^2]. \tag{2.8}$$

We wish to show that $I\geqq 0$. We divide the real axis into three intervals. Because of the symmetry in (2.8) we need only consider the following cases:

(i) $\beta_1(0)>(cF_0/2T)$, $-(cF_0/2T)\leqq\beta_2(0)\leqq(cF_0/2T)$;

(ii) $\beta_1(0)<-(cF_0/2T)$, $-(cF_0/2T)\leqq\beta_2(0)\leqq(cF_0/2T)$;

(iii) $\beta_1(0)>(cF_0/2T)$, $\beta_2(0)<-(cF_0/2T)$;

(iv) $\beta_1(0)$, $\beta_2(0)$ belong to the same interval, $\beta_1(0)>\beta_2(0)$.

Case (i). $\beta_1(0)>(cF_0/2T)$, $-(cF_0/2T)\leqq\beta_2(0)\leqq(cF_0/2T)$.

Write β_1, β_2 for $\beta_1(0), \beta_2(0)$. From (2.8),

$$\frac{2}{c}\cdot I=(\beta_2-\beta_1)^2-\left[-\beta_2-\frac{(-cF_0)+(T-ch(\cdot))\beta_1}{T+ch(\cdot)}\right], \qquad (h(\cdot)\equiv h((\alpha_1+\beta_1)(0)))$$

$$=(\beta_1-\beta_2)^2-\left[\beta_1-\beta_2-\left(\frac{2T\beta_1-cF_0}{T+ch(\cdot)}\right)\right]^2$$

$$=\frac{2T\beta_1-cF_0}{T+ch(\cdot)}\left[2(\beta_1-\beta_2)-\left(\frac{2T\beta_1-cF_0}{T+ch(\cdot)}\right)\right]$$

$$=\frac{2T\beta_1-cF_0}{T+ch(\cdot)}\left[\frac{2ch(\cdot)(\beta_1-\beta_2)+(cF_0-2\beta_2T)}{T+ch(\cdot)}\right]\geqq 0.$$

Case (ii). $\beta_1<-(cF_0/2T)$, $-(cF_0/2T)\leqq\beta_2\leqq(cF_0/2T)$.
The verification is similar to Case (i).
Case (iii). $\beta_1>(cF_0/2T)$, $\beta_2<-(cF_0/2T)$.

$$|\alpha_1|=\left|\frac{-cF_0+(T-ch(\cdot))\beta_1}{T+ch(\cdot)}\right|=\left|\beta_1-\frac{2T\beta_1-cF_0}{T+ch(\cdot)}\right|<|\beta_1|,$$

$$|\alpha_2|=\left|\frac{cF_0+(T-ch(\cdot))\beta_2}{T+ch(\cdot)}\right|=\left|\beta_2-\frac{2T\beta_2+cF_0}{T+ch(\cdot)}\right|<|\beta_2|.$$

$$\frac{2}{c}I=(\beta_1-\beta_2)^2-(\alpha_1-\alpha_2)^2\geqq(|\beta_1|+|\beta_2|)^2-(|\alpha_1|+|\alpha_2|)^2\geqq 0.$$

So far we have not utilized the assumption that g is nondecreasing. It is crucial for the verification of Case (iv) below.

Case (iv). $\beta_1(0)>\beta_2(0)$ belong to the same interval.

It is easy to see that $I=0$ if β_1 and β_2 lie in the interval $[-(cF_0/2T), (cF_0/2T)]$. So let us consider $\beta_1\geqq\beta_2>(cF_0/2T)$. We have

$$\frac{2}{c}I=(\beta_1-\beta_2)^2-\{(\beta_1-\beta_2)-[(\beta_1+\alpha_1)-(\beta_2+\alpha_2)]\}^2.$$

Hence $I\geqq 0$ if and only if

$$0\leqq(\beta_1+\alpha_1)-(\beta_2+\alpha_2)\leqq 2(\beta_1-\beta_2) \quad \text{for all } \beta_1\geqq\beta_2>\frac{cF_0}{2T}. \tag{2.9}$$

We also have

$$\beta_i+\alpha_i=\frac{2T\beta_i-cF_0}{T+ch(\cdot)}=\frac{2(\beta_i-(cF_0/2T))}{1+(c/T)h(\cdot)}, \qquad i=1,2,$$

$$(\beta_i+\alpha_i)+\frac{c}{T}g(\beta_i+\alpha_i)=2\left(\beta_i-\frac{cF_0}{2T}\right), \qquad i=1,2, \tag{2.10}$$

$$(\beta_1+\alpha_1)-(\beta_2+\alpha_2)+\frac{c}{T}[g(\beta_1+\alpha_1)-g(\beta_2+\alpha_2)]=2(\beta_1-\beta_2). \tag{2.11}$$

Sufficiency. Assume that g is nondecreasing. For fixed $\beta, \beta+\alpha$ (hence α) is determined by (2.10). Note that now the function

$$G(\eta)\equiv\eta+\frac{c}{T}g(\eta)$$

is strictly increasing because g is nondecreasing. This gives

$$(\beta_1+\alpha_1)-(\beta_2+\alpha_2)\geqq 0 \quad \text{for all } \beta_1\geqq\beta_2>\frac{cF_0}{2T}.$$

From (2.11) we get

$$(\beta_1+\alpha_1)-(\beta_2+\alpha_2)\leqq 2(\beta_1-\beta_2),$$

so (2.9) is satisfied.

Necessity. Assume that A is monotone. Then (2.9) holds. Assume the contrary, i.e., that g is not nondecreasing. Then there exist $\eta_1>\eta_2>0$ such that $g(\eta_1)<g(\eta_2)$. Let $\beta_1+\alpha_1=\eta_1$, $\beta_2+\alpha_2=\eta_2$. The values of α_i, β_i, $i=1,2$ are determined by (2.5). By (2.9), we have

$$\beta_1>\beta_2>\frac{cF_0}{2T}.$$

From (2.10), we have

$$2(\beta_1-\beta_2)=(\beta_1+\alpha_1)-(\beta_2+\alpha_2)+\frac{c}{T}[g(\beta_1+\alpha_1)-g(\beta_2+\alpha_2)]$$

$$<(\beta_1+\alpha_1)-(\beta_2+\alpha_2), \quad \text{for some } \beta_1>\beta_2>\frac{cF_0}{2T},$$

as the term in the bracket is negative. This contradicts (2.9). Therefore, g must be nondecreasing.

The case when $\beta_1<\beta_2<-(cF_0/2T)$ can be treated similarly. □

LEMMA 2. *Let $\beta\in R$ be given. Then the implicit equation*

$$\alpha=\begin{cases}\dfrac{cF_0+[T-ch(\alpha+\beta)]\beta}{T+ch(\alpha+\beta)}, & \text{if } \beta<-\dfrac{cF_0}{2T},\\[2ex] -\beta, & \text{if } -\dfrac{cF_0}{2T}\leqq\beta\leqq\dfrac{cF_0}{2T},\\[2ex] \dfrac{-cF_0+[T-ch(\alpha+\beta)]\beta}{T+ch(\alpha+\beta)}, & \text{if } \beta>\dfrac{cF_0}{2T},\end{cases}$$

has a unique solution α for any $\beta\in R$ (cf. (2.5)).

Proof. If $\beta\in[-(cF_0/2T),(cF_0/2T)]$, then $\alpha=-\beta$, so α is unique. Consider the case $\beta>(cF_0/2T)$. α satisfies

$$\alpha=\frac{-cF_0+[T-ch(\alpha+\beta)]\beta}{T+ch(\alpha+\beta)}$$

$$=\frac{[-T-ch(\alpha+\beta)]\beta+2T\beta-cF_0}{T+ch(\alpha+\beta)}$$

$$=-\beta+\frac{-cF_0+2T\beta}{T+ch(\alpha+\beta)}.$$

So

$$(\alpha+\beta)\left[1+\frac{c}{T}h(\alpha+\beta)\right]=2\beta-\frac{cF_0}{T},$$

i.e.,

$$(\alpha+\beta)+\frac{c}{T}g(\alpha+\beta)=2\beta-\frac{cF_0}{T}.$$

As noted in the proof of Lemma 1 (cf. the arguments after (2.11)), the left-hand side above is equal to $G(\alpha+\beta)$, where G is strictly increasing (as g is nondecreasing) and the range of $G(\eta)$ for $\eta \geqq 0$ is $[0, +\infty)$. Therefore the equation

$$G(\eta)=\eta+\frac{c}{T}g(\eta)=2\left(\beta-\frac{cF_0}{2T}\right)$$

has a unique solution $\eta=\alpha+\beta$ because $\beta-(cF_0/2T)>0$. Hence the solution α is unique for given $\beta>(cF_0/2T)$.

If β satisfies $\beta<-cF_0/2T$, then we have instead that α satisfies

$$(\alpha+\beta)+\frac{c}{T}g(\alpha+\beta)=2\beta+\frac{cF_0}{T}<0.$$

Again, we note that G is strictly increasing and the range of $G(\eta)$ for $\eta \leqq 0$ is $(-\infty, 0]$. Therefore α is uniquely solvable for the similar reason as above. □

Remarks.

(i) From the physical point of view, it is completely natural that g be nondecreasing, because the magnitude of frictional force increases with respect to velocity.

(ii) From Lemma 2 we can see that α is uniquely determined once β is given in (2.5). Hence, the solution of the wave equation as constructed by the method of characteristics is *unique*. This solution coincides with the unique solution constructed from the nonlinear semigroup approach.

(iii) Let $(\alpha, \beta) \in D(A)$. Then β is continuous on $[0, 1]$ so $\beta(0)$ is defined. Then $\alpha(0)$ as determined from the multivalued equation

$$\frac{T}{c}(\beta(0)-\alpha(0)) \in \bar{f}(\alpha(0)+\beta(0)) \quad \text{(cf. (1.5) or (2.5))}$$

is unique (cf. (ii) above). This implies that the operator $-A$ is single valued and coincides with $-A^0$, its minimal section.

(iv) The above discussion also shows that the solution obtained by the method of characteristics is unique as long as the function

$$G(\eta) \equiv \eta+\frac{c}{T}g(\eta)$$

is strictly increasing, i.e., g need not be nondecreasing if we are only concerned with the uniqueness of solutions by the method of characteristics.

LEMMA 3. *A is maximal monotone.*

Proof. We want to show that for any $(\phi, \psi) \in H$, there exists an $(\alpha, \beta) \in D(A)$ such that

$$(I+A)\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\begin{bmatrix}1+c\dfrac{\partial}{\partial x} & 0\\ 0 & 1-c\dfrac{\partial}{\partial x}\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\begin{bmatrix}\phi\\ \psi\end{bmatrix}. \tag{2.12}$$

We integrate (2.12) directly and obtain

$$\alpha(x) = \gamma_1 \exp((x-1)/c) + \frac{1}{c}\int_1^x \exp(-(x-\tau)/c)\phi(\tau)\,d\tau, \tag{2.13}$$

$$\beta(x) = \gamma_2 \exp((x-1)/c) - \frac{1}{c}\int_1^x \exp((x-\tau)/c)\psi(\tau)\,d\tau. \tag{2.14}$$

Condition (2.3) yields

$$\gamma_2 = -\gamma_1. \tag{2.15}$$

Thus

$$\alpha(0) = \gamma_1 e^{1/c} + \frac{1}{c}\int_1^0 e^{\tau/c}\phi(\tau)\,d\tau \equiv \gamma_1 e^{1/c} + \eta_1,$$

$$\beta(0) = -\gamma_1 e^{-1/c} - \frac{1}{c}\int_1^0 e^{-\tau/c}\psi(\tau)\,d\tau \equiv -\gamma_1 e^{-1/c} + \eta_2.$$

The constant γ_1 remains to be determined. We have three possibilities:

(i) $-(cF_0/2T) \leqq \beta(0) \leqq (cF_0/2T)$,

(ii) $\beta(0) > (cF_0/2T)$,

(iii) $\beta(0) < -(cF_0/2T)$.

Assume (i). Then $\alpha(0) = -\beta(0)$, so

$$\begin{aligned} \gamma_1 e^{1/c} + \eta_1 &= -(-\gamma_1 e^{-1/c} + \eta_2). \\ \gamma_1 &= -(\eta_1 + \eta_2)/(e^{1/c} - e^{-1/c}). \end{aligned} \tag{2.16}$$

For (i) to happen, we must have

$$-\frac{cF_0}{2T} \leqq \beta(0) = \frac{\eta_1 + \eta_2}{e^{1/c} - e^{-1/c}} \cdot e^{-1/c} + \eta_2 \leqq \frac{cF_0}{2T},$$

i.e.,

$$-\left(\frac{cF_0}{2T} + \eta_2\right)(e^{2/c} - 1) - \eta_2 \leqq \eta_1 \leqq \left(\frac{cF_0}{2T} - \eta_2\right)(e^{2/c} - 1) - \eta_2. \tag{2.17}$$

Next, assume (ii). Then by (2.5),

$$\alpha(0) = \gamma_1 e^{1/c} + \eta_1 = \frac{-cF_0 + [T - ch((\alpha+\beta)(0))](-\gamma_1 e^{-1/c} + \eta_2)}{T + ch((\alpha+\beta)(0))}. \tag{2.18}$$

By (2.4), $\eta h(\eta) = g(\eta)$. We substitute g and h and rewrite the above relation as

$$-\frac{T}{c}\{\gamma_1(e^{1/c} + e^{-1/c}) + \eta_1 - \eta_2\} = F_0 + g(\gamma_1(e^{1/c} - e^{-1/c}) + \eta_1 + \eta_2), \tag{2.19}$$

or

$$G_1(\gamma_1) = G_2(\gamma_1), \tag{2.20}$$

where G_1 and G_2 are defined, respectively, by the left- and right-hand sides of (2.19). If $g: \mathbb{R} \to \mathbb{R}$ is a bounded function, then (2.20) has a solution $\gamma_1 \in \mathbb{R}$. This is an immediate consequence of the Brouwer fixed point theorem. If g is not bounded, then by (1.7), g satisfies

$$g(\eta) \to \pm\infty \quad \text{as } \eta \to \pm\infty.$$

Thus G_1 and G_2 are continuous functions satisfying

$$G_1(\eta)\to\pm\infty \quad \text{as } \eta\to\mp\infty,$$

$$G_2(\eta)\to\pm\infty \quad \text{as } \eta\to\pm\infty,$$

so

$$G_1(\eta)-G_2(\eta)<0 \quad \text{as } \eta\to\infty,$$

$$G_1(\eta)-G_2(\eta)>0 \quad \text{as } \eta\to\infty.$$

Therefore (2.20), and hence (2.19), has a real solution γ_1.

In order for case (ii) to happen, η_1 and η_2 must satisfy

$$\beta(0)=\frac{cF_0-(T+ch((\beta+\alpha)(0))\eta_1+[T-ch((\alpha+\beta)(0))]\eta_2}{[T+ch((\alpha+\beta)(0))]e^{1/c}+[T-ch((\alpha+\beta)(0))]e^{-1/c}}+\eta_2>\frac{cF_0}{2T},$$

implying

$$\eta_1>\left(\frac{cF_0}{2T}-\eta_2\right)(e^{2/c}-1)-\eta_2. \tag{2.21}$$

Finally, we note that case (iii) can be treated in the same way as case (ii). Case (iii) happens when

$$\eta_1<-\left(\frac{cF_0}{2T}+\eta_2\right)(e^{2/c}-1)-\eta_2. \tag{2.22}$$

Therefore we see that depending on (2.17), (2.21), or (2.22), we have, respectively, (i), (ii), or (iii). In each case, γ_1 and γ_2 are solvable and (α,β) are given by (2.13), (2.14), which solves (2.9). □

3. The ω-limit set of the dynamical system. For a given solution y of a dynamical system

$$\begin{cases}\dfrac{d}{dt}y(\cdot,t)=f(y(\cdot,t)), & t>0,\\ y(\cdot,0)=y_0\end{cases} \tag{3.1}$$

in an infinite dimensional space, the ω-limit set of the solution is the intersection over $t\geqq 0$ of the closure of the orbit $\{y(x,t)\,|\,x\in[0,1]\}$. The ω-limit set of the dynamical system (3.1) is the union of the ω-limit sets of all the solutions of (3.1).

In this section, we will attempt to determine the ω-limit set of the system considered in §§ 1 and 2. Here by a "solution" we mean a solution $S(t)(\alpha_0,\beta_0)$ with $(\alpha_0,\beta_0)\in D(A^0)$.

Let

$$P\equiv\left\{(\alpha,\beta)\in D(A^0)\;\middle|\;-\frac{cF_0}{2T}\leqq f(x)\leqq\frac{cF_0}{2T},\quad \text{for } f=\alpha \text{ or } f=\beta\right\}.$$

It is easy to see that the conservation of energy

$$\left\|\begin{bmatrix}\alpha(\cdot,t)\\ \beta(\cdot,t)\end{bmatrix}\right\|_{\mathcal{H}}\equiv\left\|S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}\right\|_{\mathcal{H}}=\left\|\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}\right\|_{\mathcal{H}} \qquad t\geqq 0$$

is satisfied for any initial state $(\alpha_0, \beta_0) \in P$. Indeed, each solution $(\alpha(\cdot, t), \beta(\cdot, t))$ is periodic with period $\tau_0 = 2/c$, and $(\alpha(\cdot, t), \beta(\cdot, t)) \in P$ for any $t \geqq 0$. $S(t)$ is also *linear* on $\bar{P}$. $\bar{P}$ is a closed convex set with empty interior in $\mathcal{H}$ invariant under $S(t)$.

What happens if the initial state $(\alpha_0, \beta_0) \in D(A^0)$ is not in P? We have two cases.

Case (i).

(3.2) There exists $B > 0$ such that $h(x) \leqq T/c$ for all $x: |x| < B$.

We will show that there exists $T_0 > 0$ depending on (α_0, β_0) such that $S(t)(\alpha_0, \beta_0) \in P$ for all $t \geqq T_0$.

For any $x \in (0, 1)$, $t \geqq 0$, by the method of characteristics, we have

$$\beta(x, t+\tau_0) = \beta\left(1, t+\frac{1+x}{c}\right) = -\alpha\left(1, t+\frac{1+x}{c}\right) \qquad \left(\tau_0 = \frac{2}{c}\right)$$

$$= -\alpha\left(0, t+\frac{x}{c}\right)$$

$$= \begin{cases} [-cF_0 + T - ch(\cdot)]\beta(x, t)\}/[T + ch(\cdot)], & \text{if } \beta(x, t) < -\dfrac{cF_0}{2T}, \\ \beta(x, t), & \text{if } -\dfrac{cF_0}{2T} \leqq \beta(x, t) \leqq \dfrac{cF_0}{2T}, \\ [cF_0 - T - ch(\cdot)]\beta(x, t)\}/[T + ch(\cdot)], & \text{if } \beta(x, t) > \dfrac{cF_0}{2T}, \end{cases} \tag{3.3}$$

where $h(\cdot) \equiv h((\alpha+\beta)(0, t+x/c))$.

As we are primarily interested in the asymptotic behavior of solutions, we limit our discussion to those solutions $(\alpha(\cdot, t), \beta(\cdot, t))$ satisfying

$$|(\alpha+\beta)(0, t)| < B, \qquad t \geqq 0. \tag{3.4}$$

This is a very mild restriction as we expect that all solutions will lose the bulk of their energy and eventually satisfy (3.4), if B is not too small. See also the appendix.

We wish to prove that there exists $T_0 > 0$ such that

$$|\beta(x, T_0)| \leqq cF_0/2T, \quad \text{for all } x \in (0, 1).$$

Assume that there is some $x \in [0, 1]$ such that $\beta(x, 0) > cF_0/(2T)$. Then,

$$\beta(x, \tau_0) = \frac{cF_0}{T + ch((\alpha+\beta)(0, x/c))} - \frac{T - ch((\alpha+\beta)(0, x/c))}{T + ch((\alpha+\beta)(0, x/c))}\beta(x, 0)$$

$$< \frac{cF_0}{T + ch(\cdot)} - \frac{T - ch(\cdot)}{T + ch(\cdot)}\frac{cF_0}{2T} = \frac{cF_0}{2T},$$

where in the above, we use the shorthand $h(\cdot)$ in an obvious way from the previous expression. If $(-cF_0/(2T)) \leqq \beta(x, \tau_0) < (cF_0/(2T))$, then this x is all right. Otherwise this x satisfies

$$\beta(x, \tau_0) < -cF_0/(2T). \tag{3.5}$$

Therefore

$$
\begin{aligned}
&|\beta(x,0)|-|\beta(x,\tau_0)| \\
&\quad=\beta(x,0)-\left\{\frac{T-ch((\alpha+\beta)(0,x/c))}{T+ch((\alpha+\beta)(0,x/c))}\beta(x,0)-\frac{cF_0}{T+ch((\alpha+\beta)(0,x/c))}\right\} \\
&\quad=\frac{2h}{T+ch(\cdot)}\beta(x,0)+\frac{cF_0}{T+ch(\cdot)} \\
&\quad\geqq\frac{cF_0}{2T}, \quad \text{by (3.2), (3.4).}
\end{aligned}
\tag{3.6}
$$

Also, by (3.3)

$$
\begin{aligned}
\beta(x,2\tau_0)&=-\frac{cF_0}{T+ch((\alpha+\beta)(0,\tau_0+x/c))}-\frac{T-ch((\alpha+\beta)(0,\tau_0+x/c))}{T+ch((\alpha+\beta)(0,\tau_0+x/c))}\beta(x,\tau_0) \\
&>-\frac{cF_0}{T+ch(\cdot)}+\frac{T-ch(\cdot)}{T+ch(\cdot)}\frac{cF_0}{2T}=-\frac{cF_0}{2T}.
\end{aligned}
$$

If $(-cF_0/(2T))<\beta(x,2\tau_0)\leqq(cF_0/(2T))$ for this x, then we are done. Otherwise

$$
\beta(x,2\tau_0)>\frac{cF_0}{2T}, \quad \text{for this } x,
$$

$$
\begin{aligned}
|\beta(x,\tau_0)|-|\beta(x,2\tau_0)|&=|\beta(x,\tau_0)|-\left[\frac{T-ch((\alpha+\beta)(0,\tau_0+x/c))}{T+ch((\alpha+\beta)(0,\tau_0+x/c))}|\beta(x,\tau_0)|\right. \\
&\qquad\left.-\frac{cF_0}{T+ch((\alpha+\beta)(0,\tau_0+x/c))}\right] \\
&>\frac{cF_0}{2T}.
\end{aligned}
$$

If this process can be continued indefinitely, then

$$
|\beta(x,0)|-|\beta(x,2k\tau_0)|>2k\cdot\frac{cF_0}{2T}, \quad \text{for } k=1,2,3,\cdots,
$$

implying

$$
|\beta(x,0)|>|\beta(x,2k\tau_0)|+2k\cdot\frac{cF_0}{2T}\to\infty,
$$

which is impossible. Therefore for each $x\in[0,1]$, there is a positive integer $k(x)$ such that

$$
-cF_0/(2T)\leqq|\beta(x,k(x)\tau_0)|\leqq cF_0/(2T).
$$

As the interval $[0,1]$ is compact and $\beta(x,0)$ is continuous, we see that there exists an integer $n>0$ such that

$$
-CF_0/(2T)\leqq|\beta(x,n\tau_0)|\leqq cF_0/(2T), \quad \text{for all } x\in[0,1].
$$

Similarly,

$$
-cF/(2T)\leqq|\alpha(x,n\tau_0)|\leqq cF_0/(2T), \quad \text{for all } x\in[0,1],
$$

if n is chosen large enough.

Hence the dynamical system enters the set P at some time $t=t_0$. Once it enters P, energy no longer decays and the solution becomes periodic.

Case (ii).

(3.7) There exists $B>0$ such that $h(x)>T/c$ for all $x: |x|<B$.

Again, for the simplicity of presentation, we assume that (3.4) is in force.

For Case (ii), we will show that if the initial state (α_0, β_0) doesn't lie in P, then

$$S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}\notin P \quad \text{for all } t\geqq 0, \tag{3.8}$$

and

$$\lim_{t\to\infty} d\left(S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}, \bar{P}\right)=0, \tag{3.9}$$

where d denotes the distance metric.

Assume that $\beta(x, t)>(cF_0/(2T))$ for some $x\in[0, 1]$. Then by (3.3)

$$\begin{aligned}\beta(x, t+\tau_0)&=\frac{cF_0}{T+ch((\alpha+\beta)(0, t+x/c))}+\frac{ch((\alpha+\beta)(0, t+x/c))-T}{ch((\alpha+\beta)(0, t+x/c))+T}\beta(x, t)\\ &>\frac{cF_0}{T+ch(\cdot)}+\frac{ch(\cdot)-T}{ch(\cdot)+T}\frac{cF_0}{2T}=\frac{cF_0}{2T}.\end{aligned}$$

Therefore

$$\beta(x, t+\tau_0)>\frac{cF_0}{2T}$$

for this x, and

$$\begin{aligned}\beta(x, (k+1)\tau_0)&=\frac{cF_0}{T+ch((\alpha+\beta)(0, k\tau_0+t+x/c))}\\ &\quad+\frac{ch((\alpha+\beta)(0, k\tau_0+t+x/c))-T}{ch((\alpha+\beta)(0, k\tau_0+t+x/c))+T}\beta(x, k\tau_0)\\ &=\frac{cF_0}{T+ch(\cdot)}+\frac{ch(\cdot)-T}{ch(\cdot)+T}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]+\frac{ch(\cdot)-T}{ch(\cdot)+T}\frac{F_0}{2c}\\ &=\frac{cF_0}{2T}+\frac{ch(\cdot)-T}{ch(\cdot)+T}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right], \qquad k=0, 1, 2, \cdots.\end{aligned}$$

That is,

$$0<\beta(x, (k+1)\tau_0)-\frac{cF_0}{2T}=\frac{ch(\cdot)-T}{ch(\cdot)+T}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]. \tag{3.10}$$

If $x\in[0, 1]$ satisfies $\beta(x, t)<(-cF_0/2T)$, then we can do similarly and obtain

$$0>\beta(x, (k+1)\tau_0)+\frac{cF_0}{2T}=\frac{ch(\cdot)-T}{ch(\cdot)+T}\left[\beta(x, k\tau_0)+\frac{cF_0}{2T}\right]. \tag{3.11}$$

Identical relations (3.10) and (3.11) hold when β is replaced by α in (3.10) and (3.11). Hence the conclusion (3.8) and (3.9) follows.

From the above analysis, we can easily see that the ω-limit set of all solutions satisfying (3.4), (3.2), and (3.7) is $\bar{P}$.

4. Rate of convergence to the ω-limit set. For Case (ii) considered in the previous section, we can actually determine the rate of convergence of (3.9) if more information on the function $h(x)$ is available.

In engineering, the function g in (1.6), (1.7) is often approximated by a single polynomial function

$$g(x)=\begin{cases}\eta x^{\mu}, & x>0\\ -\eta|x|^{\mu}, & x<0\end{cases} \tag{4.1}$$

for some $\eta,\ \mu\in\mathbb{R},\ \eta,\ \mu>0$. Then

$$h(x)=\eta|x|^{\mu-1},\qquad x\in\mathbb{R},\qquad x\neq 0.$$

If $\mu>1$, then Case (i) in § 3 holds and

$$d\left(S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix},\bar{P}\right)=0\quad\text{for } t\geqq T_0\quad\text{for some } T_0\geqq 0.$$

If $\mu=1$, and $\eta\leqq T/c$, again Case (i) in § 3 is valid and (4.2) holds.

Therefore we need only consider

(a) $\mu=1$ and $\eta>T/c$; and

(b) $0<\mu<1$.

Our estimates are based on the relationships

$$\begin{aligned}&|\alpha(x,\,t+(k+1)\tau_0)|-\frac{cF_0}{2T}\\ &=\frac{ch((\alpha+\beta)(0,\,t+k\tau_0+x/c))-T}{ch((\alpha+\beta)(0,\,t+k\tau_0+x/c))+T}\left[|\alpha(x,\,t+k\tau_0)|-\frac{cF_0}{2T}\right]\end{aligned} \tag{4.2}$$

$$\begin{aligned}&|\beta(x,\,t+(k+1)\tau_0)|-\frac{cF_0}{2T}\\ &=\frac{ch((\alpha+\beta)(0,\,t+k\tau_0+x/c))-T}{ch((\alpha+\beta)(0,\,t+k\tau_0+x/c))-T}\left[|\beta(x,\,t+k\tau_0)|-\frac{cF_0}{2T}\right],\end{aligned} \tag{4.3}$$

which are consequences of (3.3).

Consider (a) first. From (4.2), (4.3) we have

$$|\alpha(x,\,t)|-\frac{cF_0}{2T}\leqq\gamma\,e^{-\delta t},$$

$$|\beta(x,\,t)|-\frac{cF_0}{2T}\leqq\gamma\,e^{-\delta t},$$

where $x\in[0,\,1]$ satisfies

$$|\alpha(x,\,0)|-\frac{cF_0}{2T}>0,$$

$$|\beta(x,\,0)|-\frac{cF_0}{2T}>0,$$

and

$$\delta\equiv\left(-\ln\left|\frac{c\eta-T}{c\eta+T}\right|\right)\Big/\tau_0$$

$$\gamma=\exp(\delta\tau_0)\cdot\sup_{\substack{0\leqq t\leqq\tau_0\\ 0\leqq y\leqq 1}}\left\{\left||\alpha(y,\,t)|-\frac{cF_0}{2T}\right|+\left||\beta(y,\,t)|-\frac{cF_0}{2T}\right|\right\}.$$

Now define a truncated generalized solution $(\bar{\alpha}, \bar{\beta})$ by

$$\bar{\phi}(x,t)=\begin{cases} -\dfrac{cF_0}{2T} & \phi(x,t)<-\dfrac{cF_0}{2T} \\ \phi(x,t) & \text{if } -\dfrac{cF_0}{2T}\leqq\phi(x,t)\leqq\dfrac{cF_0}{2T} \\ \dfrac{cF_0}{2T} & \phi(x,t)>\dfrac{cF_0}{2T}, \end{cases}$$

for $\phi=\alpha$ and $\phi=\beta$. This solution is actually some $S(t)(\bar{\alpha}_0, \bar{\beta}_0)$ with $(\bar{\alpha}_0, \bar{\beta}_0)\in\overline{D(A^0)}$, thus it is indeed a generalized solution $(\bar{\alpha}(\cdot,t), \bar{\beta}(\cdot,t))$ lying in $\bar{P}$ for all $t\geqq 0$.

$$\begin{aligned} d(t)&\equiv d\left(S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}, \bar{P}\right)\leqq d\left(S(t)\begin{bmatrix}\alpha_0\\ \beta_0\end{bmatrix}, \begin{bmatrix}\bar{\alpha}(\cdot,t)\\ \bar{\beta}(\cdot,t)\end{bmatrix}\right) \\ &=\left[\int_0^1 \{[\alpha(x,t)-\bar{\alpha}(x,t)]^2+[\beta(x,t)-\bar{\beta}(x,t)]^2\}\,dx\right]^{1/2}. \end{aligned} \tag{4.4}$$

For each $t\geqq 0$, let

$$\begin{aligned} E_1^t&=\{x\in[0,1]\,|\,\alpha(x,t)<-cF_0/(2T)\} \\ E_2^t&=\{x\in[0,1]\,|\,\alpha(x,t)>cF_0/(2T)\} \\ E_3^t&=\{x\in[0,1]\,|\,\beta(x,t)<-cF_0/(2T)\} \\ E_4^t&=\{x\in[0,1]\,|\,\beta(x,t)>cF_0/(2T)\}. \end{aligned}$$

Then

$$\begin{aligned} (4.4)&\leqq\left\{\int_{E_1^t}\left[\alpha(x,t)+\frac{cF_0}{2T}\right]^2 dx+\int_{E_2^t}\left[\alpha(x,t)-\frac{cF_0}{2T}\right]^2 dx\right. \\ &\quad\left.+\int_{E_3^t}\left[\beta(x,t)+\frac{cF_0}{2T}\right]^2 dx+\int_{E_4^t}\left[\beta(x,t)-\frac{cF_0}{2T}\right]^2 dx\right\}^{1/2} \\ &\leqq\sqrt{2}[\text{measure}\,(E_1^t\cup E_2^t\cup E_3^t\cup E_4^t)]^{1/2}\gamma\exp(-\delta t). \end{aligned}$$

Therefore *the rate of convergence is exponential.*

When $F_0=0$ in (1.6), $\bar{P}$ reduces to a single point $\{(0,0)\}$. The above exponential decay result for the linear equation is well-known. Thus the result is rather sharp.

Next, consider (b):

$$h(x)=\eta|x|^{\mu-1},\qquad \eta>0,\qquad 0<\mu<1.$$

Assume that $\beta(x,0)>(cF_0/(2T))$ for some $x\in[0,1]$. Then by (3.3)

$$\begin{aligned} \beta(x,(k+1)\tau_0)-\frac{cF_0}{2T}&=\frac{ch\left((\beta+\alpha)\left(0,k\tau_0+\dfrac{x}{c}\right)\right)-T}{ch\left((\beta+\alpha)\left(0,k\tau_0+\dfrac{x}{c}\right)\right)+T}\left[\beta(x,k\tau_0)-\frac{cF_0}{2T}\right] \\ &=\left[1-\frac{2T}{ch(\cdot)+T}\right]\left[\beta(x,k\tau_0)-\frac{cF_0}{2T}\right] \\ &\leqq\left[1-\frac{T}{ch(\cdot)}\right]\left[\beta(x,k\tau_0)-\frac{cF_0}{2T}\right] \\ &=\left[1-\frac{T}{c\eta}\left|(\beta+\alpha)\left(0,k\tau_0+\frac{x}{c}\right)\right|^{1-\mu}\right]\left[\beta(x,k\tau_0)-\frac{cF_0}{2T}\right]. \end{aligned} \tag{4.5}$$

Also from (3.3)

$$\alpha\left(0, k\tau_0+\frac{x}{c}\right)=\frac{-cF_0-\left[ch\left((\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\right)-T\right]\beta\left(0, k\tau_0+\frac{x}{c}\right)}{ch\left((\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\right)+T},$$

so

$$(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)=\frac{-cF_0+2T\cdot\beta\left(0, k\tau_0+\frac{x}{c}\right)}{ch(\cdot)+T}$$

$$=2T\left[\beta\left(0, k\tau_0+\frac{x}{c}\right)-\frac{cF_0}{2T}\right]\cdot\frac{1}{ch(\cdot)}\left[1-\frac{T}{ch(\cdot)}+\frac{T^2}{c^2h^2(\cdot)}\pm\cdots\right],$$

provided that $ch((\beta+\alpha)(0, k\tau_0+x/c))\gg T$. Continuing from the above:

$$\geqq\frac{T}{ch(\cdot)}\left[\beta\left(0, k\tau_0+\frac{x}{c}\right)-\frac{cF_0}{2T}\right].$$

Hence

$$(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\geqq\frac{T}{c\eta}\left|(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\right|^{1-\mu}\left[\beta\left(0, k\tau_0+\frac{x}{c}\right)-\frac{cF_0}{2T}\right]$$

$$\left|(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\right|^{\mu}\geqq\frac{T}{c\eta}\left[(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)-\frac{cF_0}{2T}\right]$$

$$=\frac{T}{c\eta}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]. \tag{4.6}$$

$$\left|(\beta+\alpha)\left(0, k\tau_0+\frac{x}{c}\right)\right|^{1-\mu}\geqq\left(\frac{T}{c\eta}\right)^{((1-\mu)/\mu)}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]^{((1-\mu)/\mu)}.$$

Using (4.6) in (4.5), we get

$$0<\beta(x, (k+1)\tau_0)-\frac{cF_0}{2T}\leqq\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]-\left(\frac{T}{c\eta}\right)^{1/\mu}\left[\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]^{1/\mu}.$$

Similarly, if $\beta(x, 0)<-(cF_0/2T)$ for some $x\in[0, 1]$, then

$$0<-\beta(x, (k+1)\tau_0)-\frac{cF_0}{2T}\leqq\left[-\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]$$

$$-\left(\frac{T}{c\eta}\right)^{1/\mu}\left[-\beta(x, k\tau_0)-\frac{cF_0}{2T}\right]^{1/\mu}.$$

By mathematical induction, we easily prove the convergence rate

$$\left||\beta(x, k\tau)|-\frac{cF_0}{2T}\right|\leqq\frac{d}{(1+k)^{\mu/(1-\mu)}},\quad\text{for some } D>0,\quad\text{for } k=1, 2, \cdots.$$

The same estimate also holds for $\|\alpha(x, k\tau)|-(cF_0/2T)|$. Thus we have the convergence rate $\mathcal{O}((1+t)^{-\mu/1-\mu})$ for (3.9).

Appendix. Different extensions of g.

Let g be a mapping from $[-v_0, v_0]$ into $\mathbb{R}$. Let g_1, g_2 be extensions of g:

$$\begin{cases} g_i : \mathbb{R} \to \mathbb{R}, & i = 1, 2, \\ g_i(\eta) = g(\eta) & \text{for } \eta \in [-v_0, v_0] \\ g_i \text{ satisfies (1.7)}, & i = 1, 2 \\ g_i \text{ is nondecreasing.} \end{cases}$$

As in (2.4), let

$$h_i(\eta) = g_i(\eta)/\eta, \qquad i = 1, 2, \qquad x \neq 0.$$

Let A_1, A_2 be (maximal) monotone operators in $\mathscr{H}$ defined similarly as A in § 2, except that the left end boundary condition (2.5) is now replaced by

$$\text{(A.1)} \quad \alpha_i(0) = \begin{cases} \dfrac{cF_0 + [T - ch_i((\alpha_i + \beta_i)(0))]\beta(0)}{T + ch_i((\alpha_i + \beta_i)(0))}, & \beta_i(0) < -cF_0/(2T), \\ -\beta_i(0), & -cF_0/(2T) \leqq \beta_i(0) \leqq cF_0/(2T), \\ \dfrac{-cF_0 + [T - ch_i((\alpha_i + \beta_i)(0))]\beta(0)}{T + ch_i((\alpha_i + \beta_i)(0))}, & \beta_i(0) > cF_0/(2T), \end{cases}$$

with $i = 1, 2$ for A_1 and A_2, respectively.

Assume that a given initial condition $(\alpha_0(\cdot), \beta_0(\cdot)) \in H^1(0, 1) \times H^1(0, 1)$ satisfies

$$\text{(A.2)} \quad \begin{cases} |\alpha_0(x)| \leqq \dfrac{v_0}{2} + \dfrac{cF_0}{2T} \\ |\beta_0(x)| \leqq \dfrac{v_0}{2} + \dfrac{cF_0}{2T} \end{cases}$$

on $[0, 1]$, and that $(\alpha_0, \beta_0) \in D(A_1^0)$. (The "0" superscript in $-A_1^0$ denotes the minimal section of $-A_1$.) Let $(\alpha_1(\cdot, t), \beta_1(\cdot, t)) = S_1(t)(\alpha_0, \beta_0)$ be the solution, where $S_1(t)$ is the nonlinear contraction semigroup generated by $-A_1^0$. We will show that $(\alpha_0, \beta_0) \in D(A_2^0)$ and that

$$\text{(A.3)} \qquad (\alpha_2(\cdot, t), \beta_2(\cdot, t)) = (\alpha_1(\cdot, t), \beta_1(\cdot, t)), \qquad t \geqq 0,$$

where $(\alpha_2(\cdot, t), \beta_2(\cdot, t)) = S_2(t)(\alpha_0, \beta_0)$ and $S_2(t)$ is the nonlinear contraction semigroup generated by $-A_2^0$.

First, due to (2.3), (A.1), (A.2), by the method of characteristics (cf. remarks after Lemma 2) it is easy to see that

$$|\alpha_1(x, t)| \leqq \frac{v_0}{2} + \frac{cF_0}{2T}$$

$$|\beta_1(x, t)| \leqq \frac{v_0}{2} + \frac{cF_0}{2T}$$

for all $t \geqq 0$. If $\beta_1(0, t)$ satisfies

$$|\beta_1(0, t)| \leqq cF_0/(2T),$$

then by (A.1),

$$\text{(A.4)} \qquad (\alpha_1 + \beta_1)(0, t) = 0.$$

If

$$\frac{cF_0}{2T} < \beta_1(0, t) \leqq \frac{v_0}{2} + \frac{cF_0}{2T},$$

then

$$\begin{aligned} 0 < (\alpha_1 + \beta_1)(0, t) &= \frac{-cF + 2T\beta_1(0, t)}{T + ch_1((\alpha_1 + \beta_1)(0, t))} \\ &\leqq \frac{-cF_0 + 2T\beta_1(0, t)}{T} \leqq v_0. \end{aligned} \tag{A.5}$$

Similarly, if

$$-\frac{v_0}{2} - \frac{cF_0}{2T} \leqq \beta_1(0, t) < -\frac{cF_0}{2t},$$

$$\begin{aligned} 0 > (\alpha_1 + \beta_1)(0, t) &= \frac{cF_0 + 2T\beta_1(0, t)}{T + ch_1((\alpha_1 + \beta_1)(0, t))} \\ &\geqq \frac{cF_0 + 2T\beta_1(0, t)}{T} \geqq -v_0. \end{aligned} \tag{A.6}$$

Combining (A.4), (A.5), and (A.6), we see that for all $t \geqq 0$,

$$|(\alpha_1 + \beta_1)(0, t)| \leqq v_0.$$

Since $h_1(x) = h_2(x)$ for $0 < |x| \leqq v_0$, we see that $(\alpha_0, \beta_0) \in D(A_2^0)$, and that (A.3) holds as a consequence of the method of characteristics.

Acknowledgments. The correct form of Lemma 1 was pointed out by the referees. We thank them for helpful comments.

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, the Netherlands, 1976.

[2] M. G. CRANDALL, *Lecture notes on nonlinear semigroup theory*, unpublished manuscript, University of Wisconsin, Madison, WI, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 776-785, July 1989

006

A BOUNDARY VALUE PROBLEM FOR THE MINIMUM-TIME FUNCTION*

MARTINO BARDI†

Abstract. A natural boundary value problem for the dynamic programming partial differential equation associated with the minimum time problem is proposed. The minimum time function is shown to be the unique viscosity solution of this boundary value problem.

Key words. nonlinear systems, time-optimal control, dynamic programming, Hamilton-Jacobi equations, viscosity solutions

AMS(MOS) subject classifications. 35F30, 49C20

1. Introduction.

Given the control system

$$\dot{y}(s)+b(y(s), z(s))=0, \qquad s\geqq 0, \quad y\in\mathbb{R}^N, \quad z\in Z\subseteq\mathbb{R}^M, \tag{1.1}$$

the minimum time function $T(x)$ associates to a point $x\in\mathbb{R}^N$ the infimum of the times that the trajectories of (1.1) satisfying $y(0)=x$ take to reach the origin. The problem of determining T and the optimal controls realizing the minimum is one of the most extensively studied in the control-theoretic literature, especially in the linear case [1], [5], [6], [8], [15], [16], [20], [22]-[26].

It is well known that Bellman's Dynamic Programming Principle implies that at points of differentiability T satisfies the following first-order fully nonlinear partial differential equation (PDE) of Hamilton-Jacobi type:

$$\sup_{z\in Z} b(x, z)\cdot DT(x)=1. \tag{1.2}$$

It is also known that in general T is not differentiable everywhere, but it satisfies (1.2) in some generalized sense [8], [24]. In the last five years the new concept of viscosity solution for Hamilton-Jacobi equations has been introduced by Crandall and Lions [11] and the theory of such solutions has developed quickly (see, e.g., [9]-[12], [17], [21] and the references therein) and has been applied to many problems in control theory and differential games (see, e.g., [3], [4], [7], [13], [21], the survey paper of Fleming [14], and its long list of references). Following Lions [21] it is not hard to show that T satisfies (1.2) in the viscosity sense as soon as it is continuous. The goal of this paper is to complement (1.2) with a natural boundary condition and prove a uniqueness result for viscosity solutions of such a boundary value problem (BVP). Ishii [17] has proved the uniqueness of viscosity solutions of the Dirichlet problem in a bounded open set for a class of equations that includes (1.2). However the Dirichlet problem does not seem to be the most natural one for the minimum time function because in general we do not know a priori the value of T on the boundary of some given bounded set. Instead we consider (1.2) in the set $\mathcal{R}\backslash\{0\}$ where $\mathcal{R}$ is the set of points x such that there is a trajectory of (1.1) starting at x and reaching the origin in finite time, i.e., the largest set where T is defined (and finite); we propose the following boundary condition:

$$T(0)=0 \quad \text{and} \quad T(x)\to+\infty \quad \text{uniformly as } x\to\partial\mathcal{R}. \tag{1.3}$$

In § 2 we will prove that under quite general assumptions T satisfies (1.2) in $\mathcal{R}\backslash\{0\}$ in the viscosity sense and (1.3). In § 3 we will show that for any open set $\mathcal{R}$ having

* Received by the editors December 9, 1987; accepted for publication (in revised form) July 7, 1988.
† Dipartimento di Matematica Pura e Applicata, Università di Padova, I-35131 Padova, Italy.

zero in its interior there is at most one viscosity solution of (1.2) in $\mathcal{R}\backslash\{0\}$ satisfying (1.3) and bounded below.

The main result is the uniqueness theorem in § 3. It presents three difficulties with respect to standard uniqueness results in the Hamilton-Jacobi theory: (i) the Hamiltonian depends on the gradient of the unknown function but does not depend explicitly on the unknown function itself, while it is usually required that it be strictly monotone in such a variable; (ii) the infinite boundary condition (1.3) if $\mathcal{R} \neq \mathbb{R}^N$; and (iii) the lack of regularity of the solutions to be compared and of the Hamiltonian if b is not globally bounded and $\mathcal{R}$ is unbounded.

To overcome the first difficulty we introduce a change of the unknown variable first used for this goal by Kruzkov [18]. It turns out that this transforms the infinite boundary condition into a finite one, therefore automatically taking care of the second difficulty.

The third difficulty can be solved using the new approach to uniqueness presented in the paper of Crandall, Ishii, and Lions [10]. Indeed our uniqueness theorem has to be considered as a corollary of the methods developed in [10]. As one of the referees pointed out to us, this difficulty had already been overcome for different problems in an earlier paper by Ishii [28], whose methods could be applied effectively to our problem as well.

We remark that the proof of the uniqueness theorem does not make use of the convexity of the Hamiltonian. Therefore the methods of this paper can be employed to study the minimum time problem in games of pursuit and evasion (see Bardi and Soravia [27]).

After the completion of this work we have learned that Kruzkov's change of variables has been used recently by Lasry and Lions [19] to study the minimum time function of a differential game with state constraints in a bounded domain, and by Barles [2] for unbounded control problems. Moreover, one of the referees pointed out that a uniqueness theorem for discontinuous solutions of (1.2), (1.3) can be proved by combining the Kruzkov transform and the results by Barles and Perthame [3], provided the target to be reached is a smooth set instead of a single point.

In the last decade the use of the theory of subanalytic sets has led to very strong results on the regularity of the minimum time function and of feedback controls (see Brunovsky [6] and Sussmann [26] and the references therein). However it is also known that certain quite smooth systems exhibit very irregular behaviors that fail to fall within the theory of subanalyticity (see Lojasiewicz and Sussmann [22] or the classical Fuller's example in [23]). It is our hope that the PDE setting of the minimum time problem proposed in this paper could be of some help in the study of these problems.

2. The BVP of the minimum time function. We begin listing the assumptions to be used in the following.

(H1) $b:\mathbb{R}^N \times Z \to \mathbb{R}^N$, where $Z \subseteq \mathbb{R}^M$, is continuous and there exist L, $K \in \mathbb{R}$ such that $|b(x,z)-b(y,z)| \leqq L|x-y|$ and $|b(y,z)| \leqq K(1+|y|)$, for all x, $y \in \mathbb{R}^N$, and for all $z \in Z$.

Let $\mathcal{M}$ be the set of measurable functions $z:[0,\infty) \to Z$, and let $y(s)=y(s;x,z)$ be the solution of

$$y(s) = x - \int_0^s b(y(t), z(t))\, dt \tag{2.0}$$

for $s \geqq 0$, $x \in \mathbb{R}^N$, and $z \in \mathcal{M}$. Let $\mathcal{T} \subseteq \mathbb{R}^N$ be a given closed set, the *terminal set* (e.g.,

$\mathcal{T} = \{0\}$), and define

$$\mathcal{R} := \{x \in \mathbb{R}^N : \exists z \in \mathcal{M},\ t \geqq 0 \text{ such that } y(t; x, z) \in \mathcal{T}\}.$$

Clearly, $\mathcal{R} \supseteq \mathcal{T}$. Now define the minimum time function

$$T : \mathcal{R} \to [0, \infty), \qquad T(x) := \inf\{t : y(t; x, z) \in \mathcal{T} \text{ for some } z \in \mathcal{M}\}.$$

We will assume the following.

(H2) $\mathcal{R}$ is open.

(H3) T is continuous in $\mathcal{R}$.

(H4) For every $x_0 \in \partial\mathcal{R}$, $\lim_{x \to x_0} T(x) = +\infty$.

Conditions under which (H2)-(H4) are satisfied are well known in the literature, especially in the linear case. See, for instance, [1], [15], [20] for (H2), [25], [15], [1], [5], [8] for (H3), [8], [15] for (H4), and the references therein. Essentially (H2)-(H4) follow from some controllability around $\mathcal{T}$.

Next we define a suitable boundary condition for functions $u \in C(\mathcal{R} \setminus \mathcal{T})$:

(BC) u converges uniformly to 0 as $x \to \partial\mathcal{T}$ and to $+\infty$ as $x \to \partial\mathcal{R}$, i.e., for every ε, $M > 0$ there exists $\delta > 0$ such that $\operatorname{dist}(x, \partial\mathcal{T}) < \delta$ implies $|u(x)| < \varepsilon$ and $\operatorname{dist}(x, \partial\mathcal{R}) < \delta$ implies $u(x) > M$.

Before proving that T satisfies (BC) we need a technical lemma.

LEMMA 1. *Assume* (H1) *is true. If* $y(t; x, z) = x_1$, *then*

$$t \geqq \frac{1}{K} \log\left(1 + \frac{|x - x_1|}{1 + \min(|x|, |x_1|)}\right).$$

Proof. Hypothesis (H1) implies

$$|y(s) - x| \leqq Ks(1 + |x|) + \int_0^s K|y(\tau) - x|\, d\tau,$$

and then, using Gronwall's inequality, we get

$$|x_1 - x| \leqq (1 + |x|)(e^{Kt} - 1),$$

which gives

$$t \geqq \frac{1}{K} \log\left(1 + \frac{|x_1 - x|}{1 + |x|}\right).$$

The same calculation for $\tilde{y}(s) := y(t - s)$ leads to

$$t \geqq \frac{1}{K} \log\left(1 + \frac{|x_1 - x|}{1 + |x_1|}\right). \qquad \square$$

Remark 1. It follows easily from Lemma 1, choosing $x_1 \in \mathcal{T}$, that $T(x) > 0$ for all $x \in \mathcal{R} \setminus \mathcal{T}$ (since $\mathcal{T}$ is closed) and that $\mathcal{T}$ bounded implies $\lim_{|x| \to \infty} T(x) = +\infty$.

LEMMA 2. *Assume* (H1)-(H4) *and* $\mathcal{T}$ *bounded. Then* T *satisfies* (BC).

Proof. Since T is continuous, null on $\partial\mathcal{T}$, and $\mathcal{T}$ is bounded, the first part of (BC) is clearly satisfied.

To prove the second part we fix $M > 0$. From Lemma 1 and Remark 1 the existence of $R > 0$ such that $T(x) > M$ for $|x| > R$ follows. Now we use (H4) to get a covering of the compact set $\partial\mathcal{R} \cap \{x : |x| \leqq R\}$ made of a finite number of open balls B_i centered on $\partial\mathcal{R}$ and having small radius so that $T(x) > M$ for $x \in \mathcal{R} \cap B_i$. We conclude observing that there exists $\delta > 0$ such that $|x| \leqq R$ and $\operatorname{dist}(x, \partial\mathcal{R}) < \delta$ imply $x \in B_i$ for some i. $\square$

We recall that a continuous function u defined in an open set $\mathcal{O} \subseteq \mathbb{R}^N$ is defined to be a *viscosity solution* of

$$H(x, u, Du) = 0 \quad \text{in } \mathcal{O},$$

if for every $\phi \in C^1(\mathcal{O})$ and x_0 local maximum point of $u - \phi$ we have

$$H(x_0, u(x_0), D\phi(x_0)) \leqq 0,$$

while for x_0, local minimum point of $u - \phi$ we have

$$H(x_0, u(x_0), D\phi(x_0)) \geqq 0.$$

We will now prove that T is a viscosity solution of

$$\sup_{z \in Z} b(x, z) \cdot Du - 1 = 0 \quad \text{in } \mathcal{R} \setminus \mathcal{T}. \tag{HJ}$$

This fact is certainly known to experts, since it follows from arguments of Lions [21, Chap. 1]. We include a full proof for the sake of completeness.

Define for $z \in \mathcal{M}$, $x \in \mathbb{R}^N$, $t_x(z) := \inf\{t: y(t; x, z) \in \mathcal{T}\}$ and denote by $\chi_{\{t < t_x(z)\}}$ the function defined on $[0, \infty)$ which is one if $t < t_x(z)$ and zero if the opposite inequality holds.

LEMMA 3 (Dynamic Programming Principle). *Assume* (H1). *Then for all* $x \in \mathcal{R}$ *and* $t \geqq 0$

$$T(x) = \inf_{z \in \mathcal{M}} \{\min(t, t_x(z)) + \chi_{\{t < t_x(z)\}} T(y(t; x, z))\}.$$

Proof. Fix x and t and let A be the right-hand side of the above equality. To prove $T(x) \geqq A$ we fix an arbitrary $\varepsilon > 0$ and show that

$$T(x) \geqq A - \varepsilon. \tag{2.1}$$

Let $z_1 \in \mathcal{M}$ be such that

$$T(x) \geqq t_x(z_1) - \varepsilon. \tag{2.2}$$

If $t \geqq t_x(z_1)$, then (2.1) holds. Now suppose $t < t_x(z_1)$ and let $z_2(s) = z_1(t + s)$. Then

$$t_x(z_1) = t + t_{y(t;x,z_1)}(z_2) \geqq t + T(y(t; x, z_1)) \geqq A,$$

and so by (2.2) we have (2.1).

Now we want to prove

$$T(x) \leqq A + \varepsilon, \tag{2.3}$$

and for this we choose z_1 such that

$$A + \frac{\varepsilon}{2} \geqq \min(t, t_x(z_1)) + \chi_{\{t < t_x(z_1)\}} T(y(t; x, z_1)). \tag{2.4}$$

If $t \geqq t_x(z_1)$ then (2.3) holds. If $t < t_x(z_1)$ let $z_2 \in \mathcal{M}$ be such that

$$T(y(t; x, z_1)) \geqq t_{y(t;x,z_1)}(z_2) - \frac{\varepsilon}{2}. \tag{2.5}$$

Now define the control

$$z_3(s) := \begin{cases} z_1(s) & \text{if } s < t, \\ z_2(s - t) & \text{if } s \geqq t. \end{cases}$$

Clearly

$$t_x(z_3) = t + t_{y(t;x,z_1)}(z_2),$$

and therefore we get from (2.4) and (2.5)

$$A + \frac{\varepsilon}{2} \geqq t + T(y(t; x_1, z_1)) \geqq t_x(z_3) - \frac{\varepsilon}{2} \geqq T(x) - \frac{\varepsilon}{2},$$

which proves (2.3). By the arbitrariness of ε the proof is complete. □

THEOREM 1. *Assume* (H1)-(H3). *Then the minimum time function T is a positive viscosity solution of* (HJ). *If moreover* (H4) *holds and the terminal set $\mathcal{T}$ is bounded, then T satisfies the boundary condition* (BC) *as well.*

Proof. The second statement is just Lemma 2. To prove the first statement we begin by considering $x_0 \in \mathcal{R} \setminus \mathcal{T}$ and $\phi \in C^1(\mathcal{R} \setminus \mathcal{T})$ such that for all x sufficiently close to x_0,

$$T(x_0) - \phi(x_0) \geqq T(x) - \phi(x). \tag{2.6}$$

Let z_1 be any constant control, $z_1(s) \equiv \bar{z} \in Z$. By Lemma 3 we have for all $0 < s < T(x_0)$

$$T(x_0) - T(y(s; x_0, z_1)) \leqq s,$$

and so by (2.6) we have for sufficiently small positive s,

$$\frac{\phi(x_0) - \phi(y(s; x_0, z_1))}{s} \leqq \frac{T(x_0) - T(y(s; x_0, z_1))}{s} \leqq 1.$$

Now letting $s \searrow 0$ and using (2.0), we get

$$D\phi(x_0) \cdot b(x_0, \bar{z}) \leqq 1,$$

and by the arbitrariness of $\bar{z}$,

$$\sup_{z \in Z} b(x_0, z) \cdot D\phi(x_0) - 1 \leqq 0.$$

Let us now consider new x_0 and ϕ as above but such that

$$T(x_0) - \phi(x_0) \leqq T(x) - \phi(x)$$

for all x in a given neighborhood of x_0. By Lemma 1 there exists s_1 such that

$$\phi(x_0) - \phi(y(s; x_0, z_1)) \geqq T(x_0) - T(y(s; x_0, z)) \quad \forall s \leqq s_1, \quad \forall z \in \mathcal{M}.$$

Fix $\varepsilon > 0$. By Lemma 3 for every $s \leqq T(x_0)$ there is $z^* \in \mathcal{M}$ such that

$$T(x_0) \geqq s + T(y(s; x_0, z^*)) - \varepsilon s,$$

and thus for $0 < s \leqq s_2$

$$\frac{\phi(x_0) - \phi(y(s; x_0, z^*))}{s} \geqq 1 - \varepsilon. \tag{2.7}$$

Using (2.0) and the expansion

$$\phi(x) = \phi(x_0) + D\phi(x_0) \cdot (x - x_0) + m(x)|x - x_0| \quad \text{with } \lim_{x \to x_0} m(x) = 0,$$

we can write the left-hand side of (2.7) as follows:

$$\frac{1}{s} \int_0^s D\phi(x_0) \cdot b(y(t; x_0, z^*), z^*(t))\, dt - m(y(s; x_0, z^*)) \frac{1}{s} \left| \int_0^s b(y(t; x_0, z^*), z^*(t))\, dt \right|.$$

The second term in this expression can be made smaller than ε for small s by (H1) and Lemma 1; by using (H1) again, the first term can be written as $1/s \int_0^s D\phi(x_0) \cdot b(x_0, z^*(t))\, dt$ plus a correction term smaller than ε for small s. Then by (2.7),

$$\sup_{z \in Z} D\phi(x_0) \cdot b(x_0, z) \geqq \frac{1}{s} \int_0^s D\phi(x_0) \cdot b(x_0, z^*(t))\, dt \geqq 1 - 3\varepsilon,$$

which provides the desired inequality by the arbitrariness of ε. □

3. Uniqueness. In this section we will prove the following theorem.

THEOREM 2. *Assume* (H1), *let $\mathcal{R}$ be an open subset of $\mathbb{R}^N$, and let $\mathcal{T} \subseteq \mathcal{R}$ be a closed set. If $u_1, u_2 \in C(\mathcal{R} \setminus \mathcal{T})$ are viscosity solutions of* (HJ) *satisfying* (BC) *and bounded from below, then $u_1 = u_2$.*

The main tool of the proof is a lemma of Crandall, Ishii, and Lions [10, Lemma 1], which we report below in a version simplified for the present purpose. We say that $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfies condition (H) if the following holds:

(H) H is continuous; there exist a Lipschitz continuous, everywhere differentiable function $\mu : \mathbb{R}^n \to [0, \infty)$ and a continuous, nondecreasing function $\sigma : [0, \infty) \to [0, \infty)$ satisfying $\sigma(0) = 0$, such that $H(x, p) - H(x, p + \lambda D\mu(x)) \leqq \sigma(\lambda)$, for all x, $p \in \mathbb{R}^n$, $\lambda \in [0, 1]$ and $\lim_{|x| \to \infty} \mu(x) = +\infty$.

LEMMA 4. *Let H satisfy condition* (H) *and let Ω be an open subset of $\mathbb{R}^n$. Let $z \in C(\bar{\Omega})$ be a viscosity solution of $z + H(x, Dz) = 0$ in Ω, and let $w \in C^1(\bar{\Omega})$ satisfy*

$$w(x) + H(x, Dw(x)) \geqq 0 \quad \text{and} \quad |Dw(x)| \leqq C \quad \forall x \in \Omega.$$

Assume that

$$\sup_{\partial\Omega} (z - w) < \sup_{\Omega} (z - w) < \infty.$$

Then $z \leqq w$ in Ω.

Proof. See [10] for the proof of this lemma. □

The other key tool in the proof of Theorem 2 is a change of the unknown variable in (HJ). For this we need a slight extension of Corollary I.8 in [11].

LEMMA 5. *Let u be a viscosity solution of $H(x, u, Du) = 0$ in $\mathcal{O}$, open subset of $\mathbb{R}^n$; let $\Phi \in C^1(\mathbb{R})$, $\Phi'(r) > 0$ for all r; and let $\Psi : \Phi(\mathbb{R}) \to \mathbb{R}$ be the inverse function of Φ. Then $v = \Phi \circ u$ is a viscosity solution of*

$$H(x, \Psi(v), \Psi'(v) Dv) = 0 \quad \text{in } \mathcal{O}.$$

Proof. Let $x_0 \in \mathcal{O}$ and $\zeta \in C^1(\mathcal{O})$ be such that $v - \zeta$ has a local maximum in x_0. Define $\xi(x) = \zeta(x) - \zeta(x_0) + v(x_0)$. We have

$$D\xi = D\zeta, \qquad \xi(x_0) = v(x_0) = \Phi(u(x_0)),$$

$$v(x) \leqq \xi(x) \quad \text{in a neighborhood of } x_0.$$

Since $\Phi(\mathbb{R})$ is open, $\Psi \circ \xi$ is defined in a neighborhood of x_0 and we extend it to $\eta \in C^1(\mathcal{O})$. By the monotonicity of Ψ we have

$$u(x_0) = \eta(x_0), \quad u(x) \leqq \eta(x) \quad \text{in a neighborhood of } x_0.$$

Then

$$H(x_0, u(x_0), D\eta(x_0)) \leqq 0,$$

and so

$$H(x_0, \Psi(v(x_0)), \Psi'(v(x_0))D\zeta(x_0)) \leqq 0.$$

If x_0 is a minimum point we get the desired inequality in the same way. □

Proof of Theorem 2. Define $\Phi(t) := 1 - e^{-t}$, $v_1 := \Phi \circ u_1$, $v_2 := \Phi \circ u_2$, $\mathcal{O} := \mathcal{R} \setminus \mathcal{T}$. By Lemma 5, v_1 and v_2 are viscosity solutions of

$$\sup_{z \in Z} \left\{ b(x, z) \cdot \left(\frac{1}{1-v} Dv \right) \right\} - 1 = 0 \quad \text{in } \mathcal{O},$$

and since $v_1, v_2 < 1$, they are also viscosity solutions of

$$v + \sup_{z \in Z} \{b(x, z) \cdot Dv\} - 1 = 0 \quad \text{in } \mathcal{O}, \tag{3.0}$$

as is easy to verify from the definition. They can be extended in a unique way to v_1, $v_2 \in C(\bar{\mathcal{O}})$ by setting

$$v_i = 0 \quad \text{on } \partial\mathcal{T}, \quad v_i = 1 \quad \text{on } \partial\mathcal{R}, \quad i = 1, 2.$$

Now define

$$H(x, p) := \sup_{z \in Z} \{b(x, z) \cdot p - 1\}.$$

We claim that H satisfies condition (H). Fix $\varepsilon > 0$ and choose $z_1 \in Z$ such that $H(x, p) \leqq b(x, z_1) \cdot p - 1 + \varepsilon$. Then by (H1)

$$\begin{aligned} H(x, p) - H(y, q) &\leqq b(x, z_1) \cdot p - b(y, z_1) \cdot q + \varepsilon \\ &\leqq L|x - y||p| + K(1 + |y|)|p - q| + \varepsilon, \end{aligned}$$

and thus

$$|H(x, p) - H(y, q)| \leqq L|x - y||p| + K(1 + |y|)|p - q|, \tag{3.1}$$

which implies the continuity of H. Now let $h \in C^1(\mathbb{R})$ be such that $h(0) = h'(0) = 0$, $h(e) = 1$, $h'(e) = 1/e$, and define

$$\mu(x) := \begin{cases} h(|x|) & \text{if } |x| < e, \\ \log(|x|) & \text{if } |x| \geqq e. \end{cases}$$

Clearly, $\mu \in C^1(\mathbb{R}^N)$ and it is Lipschitz continuous since $D\mu(x) = (1/|x|^2)x$ for $|x| \geqq e$. Moreover, by (3.1),

$$H(x, p) - H(x, p + \lambda D\mu(x)) \leqq K(1 + |x|)\lambda |D\mu(x)| \leqq \lambda C =: \sigma(\lambda)$$

for $\lambda > 0$, where $C := \max(2K, K(1 + e) \sup |h'|)$, which proves the claim.

Next we define, following Crandall, Ishii, and Lions [10],

$$\hat{H}(x, y, p, q) := H(x, p) - H(y, -q),$$

$$z(x, y) := v_1(x) - v_2(y),$$

and note that z is a viscosity solution of

$$z + \hat{H}(x, y, D_x z, D_y z) = 0 \quad \text{in } \mathcal{O} \times \mathcal{O}.$$

Our goal is to prove that $z(x, x) \leqq 0$, because by interchanging the roles of v_1 and v_2 we get $v_1 = v_2$ and then $u_1 = u_2$. To reach our goal we are going to apply Lemma 4 to z defined above, $\Omega := \Delta \cap (\mathcal{O} \times \mathcal{O})$ where

$$\Delta := \{(x, y) \in \mathbb{R}^{2N} : |x - y| < 1\},$$

and $w = w_\varepsilon$ for suitable ε where

$$w_\varepsilon(x, y) := (\varepsilon^{4L} + |x-y|^2)^{1/2L}/\varepsilon.$$

We have to show that there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon \leqq \varepsilon_0$, w_ε satisfies the hypotheses of Lemma 4. Once this is done we have $z(x, x) \leqq w_\varepsilon(x, x) = \varepsilon$, and letting $\varepsilon \searrow 0$ we conclude.

Since

$$D_x w_\varepsilon(x, y) = \frac{1}{\varepsilon L}(\varepsilon^{4L} + |x-y|^2)^{(1/2L)-1}(x-y) = -D_y w_\varepsilon(x, y),$$

w_ε is Lipschitz continuous in $\bar{\Delta}$ and, moreover, by (3.1),

$$\begin{aligned} w_\varepsilon(x, y) + \hat{H}(x, y, D_x w_\varepsilon, D_y w_\varepsilon) &\geqq w_\varepsilon - L|x-y|^2(\varepsilon^{4L} + |x-y|^2)^{(1/2L)-1}/\varepsilon L \\ &\quad - K(1+|y|)|D_x w_\varepsilon + D_y w_\varepsilon| \\ &\geqq w_\varepsilon - w_\varepsilon = 0. \end{aligned}$$

Since u_1 and u_2 are bounded from below, v_1 and v_2 are bounded, and

$$\alpha_\varepsilon := \sup_\Omega (z - w_\varepsilon) \leqq A < \infty \quad \text{for all } \varepsilon > 0. \tag{3.2}$$

If

$$\liminf_{\varepsilon \searrow 0} \alpha_\varepsilon \leqq 0,$$

we immediately obtain $z(x, x) \leqq 0$. Thus it remains to prove that $\alpha_\varepsilon \geqq \alpha > 0$ and $0 < \varepsilon \leqq \varepsilon_0$ imply

$$\sup_{\partial\Omega} (z - w_\varepsilon) < \alpha_\varepsilon = \sup_\Omega (z - w_\varepsilon).$$

Suppose that $(\bar{x}, \bar{y}) \in \partial\Omega$ is such that

$$z(\bar{x}, \bar{y}) - w_\varepsilon(\bar{x}, \bar{y}) \geqq \frac{\alpha}{2} > 0. \tag{3.3}$$

Fix $0 < \delta < 1$ such that $x \in \partial\mathcal{O}$ and $|x-y| < \delta$ implies $|v_i(x) - v_i(y)| < \alpha/2$, $i = 1, 2$. This can be done because v_1 and v_2 take up their boundary values uniformly as a consequence of (BC). Suppose first that $|\bar{x} - \bar{y}| < \delta$ so that either $\bar{x}$ or $\bar{y}$, say $\bar{x}$, belongs to $\partial\mathcal{O}$. Then $z(\bar{x}, \bar{x}) = 0$ and $w_\varepsilon \geqq 0$ imply

$$z(\bar{x}, \bar{y}) - w_\varepsilon(\bar{x}, \bar{y}) \leqq v_1(\bar{x}) - v_2(\bar{y}) - v_1(\bar{x}) + v_2(\bar{x}) < \frac{\alpha}{2},$$

a contradiction to (3.3). On the other hand, if $|\bar{x} - \bar{y}| \geqq \delta$, (3.2) and (3.3) imply

$$z(\bar{x}, \bar{y}) - w_\varepsilon(\bar{x}, \bar{y}) \geqq \sup_\Omega (z - w_\varepsilon) - \alpha_\varepsilon \geqq z(\bar{x}, \bar{x}) - w_\varepsilon(\bar{x}, \bar{x}) - A,$$

from which we obtain

$$v_2(\bar{x}) - v_2(\bar{y}) \geqq w_\varepsilon(\bar{x}, \bar{y}) - \varepsilon - A.$$

The right-hand side of the last inequality can be made arbitrarily large choosing ε small because

$$\liminf_{\varepsilon \searrow 0} \{w_\varepsilon(x, y) : |x-y| \geqq \delta\} = +\infty,$$

and we get a contradiction because the left-hand side is bounded. □

Remark 2. Under the stronger assumption that $|b(y, z)| \leqq K$ for all $y \in \mathcal{R} \setminus \mathcal{T}$, $z \in Z$ (which excludes the linear case if $\mathcal{R}$ is unbounded), and strengthening the boundary conditions, we can give a weaker uniqueness theorem with a much shorter proof based on the earliest uniqueness result for viscosity solutions, that is, Theorem III.1 of Crandall and Lions [11]. In fact, under such an assumption on b, (3.1) implies that the Hamiltonian $H(x, p)$ is uniformly continuous in $\mathbb{R}^N \times \{p: |p| \leqq R\}$ for all $R > 0$. The additional boundary condition is

$$\text{(ABC)} \qquad \lim_{|x| \to \infty} u(x) = +\infty,$$

which is satisfied by the minimum time function T if $\mathcal{T}$ is bounded (see Remark 1). The proof of the weaker theorem begins with the same change of variables as the proof of Theorem 2. Next, we note that if u_1, u_2 satisfy (BC) and (ABC) then v_1, $v_2 \in BUC(\bar{\mathcal{O}})$ and that (3.0) satisfies the hypotheses of Theorem III.1(ii) in [11], which implies $v_1 = v_2$.

Remark 3. If $\mathcal{R}$ is bounded, u_1 and u_2 satisfy (ABC), they are zero on $\partial \mathcal{T}$, and they tend to $+\infty$ as $x \to x_0 \in \partial \mathcal{R}$, then the hypotheses of Theorem 2 are satisfied. In fact, u_1 and u_2 are clearly bounded from below, and they satisfy (BC) by the proof of Lemma 2.

Remark 4. If we drop the assumption that u_1 and u_2 are bounded below, then the conclusion of the theorem is false. In fact, if we take $b(x, z) = z$, Z the unit ball in $\mathbb{R}^N$, $\mathcal{T} = \{0\}$, then we get the BVP

$$|Du| - 1 = 0 \quad \text{in } \mathbb{R}^N \setminus \{0\},$$
$$u(0) = 0,$$

which has the classical solutions $u_1(x) = |x|$ and $u_2(x) = -|x|$.

Acknowledgment. The author thanks Alberto Bressan for some interesting conversations which stimulated this research.

REFERENCES

[1] A. BACCIOTTI, *Sulla continuità della funzione tempo minimo*, Boll. Un. Mat. Ital. B (6), 15 (1978), pp. 859-868.

[2] G. BARLES, *An approach of deterministic unbounded control problems and of first-order Hamilton-Jacobi equations with gradient constraints*, preprint.

[3] G. BARLES AND B. PERTHAME, *Discontinuous solutions of deterministic optimal stopping time problems*, Modélisation Mathématique et Analyse Numérique, 21 (1987), pp. 557-579.

[4] E. N. BARRON AND R. JENSEN, *The Pontryagin maximum principle from dynamic programming and viscosity solutions to first-order* PDE, Trans. Amer. Math. Soc., 298 (1986), pp. 635-641.

[5] A. BRESSAN, *Sulla funzione tempo minimo nei sistemi non lineari*, Atti. Accad. Naz. Lincei, Cl. Sci. Fis. Mat. Natur. LXVI (1979), pp. 383-388.

[6] P. BRUNOVSKY, *On the structure of optimal feedback systems*, in Proc. Internat. Congress of Mathematicians, Helsinky, 1978.

[7] I. CAPUZZO-DOLCETTA AND L. C. EVANS, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143-161.

[8] R. CONTI, *Processi di controllo lineari in* $\mathbb{R}^n$, Quad. Unione Mat. Italiana 30, Pitagore, Bologna, 1985.

[9] M. 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.

[10] M. CRANDALL, H. ISHII, AND P. L. LIONS, *Uniqueness of viscosity solutions revisited*, J. Math. Soc. Japan, 39 (1987), pp. 581-596.

[11] M. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[12a] ———, *Hamilton-Jacobi equations in infinite dimensions* I, J. Funct. Anal., 62 (1985), pp. 379-396.

[12b] ———, *Hamilton-Jacobi equations in infinite dimensions* II, J. Funct. Anal., 65 (1986), pp. 368-405.

[12c] M. CRANDALL AND P. L. LIONS, *Hamilton-Jacobi equations in infinite dimensions* III, J. Funct. Anal., 68 (1986), pp. 214-247.

[13] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 773-797.

[14] W. H. FLEMING, *Controlled Markov Processes and Viscosity Solutions of Nonlinear Evolution Equations*, Lecture Notes, Scuola Normale Superiore Pisa, 1986.

[15] O. HAJEK, *Geometric theory of time-optimal control*, SIAM J. Control Optim., 9 (1971), pp. 341-350.

[16] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[17] H. ISHII, *A simple, direct proof of uniqueness for solutions of the Hamilton-Jacobi equations of eikonal type*, Proc. Amer. Math. Soc., 100, 1987, pp. 247-251.

[18] S. N. KRUZKOV, *Generalized solutions of the Hamilton-Jacobi equations of eikonal type* I, Math. USSR Sb., 27 (1975), pp. 406-445.

[19] J. M LASRY AND P. L. LIONS, personal communication.

[20] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1968.

[21] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[22] S. LOJASIEWICZ AND H. J. SUSSMANN, *Examples of reachable sets and optimal cost functions which fail to be subanalytic*, SIAM J. Control Optim., 23 (1985), pp. 584-598.

[23] C. MARCHAL, *Chattering arcs and chattering controls*, J. Optim. Theory Appl., 11 (1973), pp. 441-468.

[24] F. MIGNANEGO AND G. PIERI, *On a generalized Bellman equation for the optimal-time problem*, Systems Control Lett., 3 (1983), pp. 235-241.

[25] N. N. PETROV, *The continuity of Bellman's generalized function*, Differential Equations, 6 (1970), pp. 290-292.

[26] H. J. SUSSMANN, *Analytic stratifications and control theory*, in Proc. Internat. Congress of Mathematicians, Helsinky, 1978.

[27] M. BARDI AND P. SORAVIA, *A* PDE *framework for games of pursuit-evasion type*, in Differential Games and Applications, T. Basar and P. Bernhard, eds., Lecture Notes in Control and Information Sci., Springer-Verlag, Berlin, New York, to appear.

[28] H. ISHII, *Uniqueness of unbounded viscosity solution of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 33 (1984), pp. 721-748.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 786-813, July 1989

007

OPTIMAL SENSOR SCHEDULING IN NONLINEAR FILTERING OF DIFFUSION PROCESSES*

JOHN S. BARAS† AND ALAIN BENSOUSSAN‡

Abstract. The nonlinear filtering problem of a vector diffusion process is considered when several noisy vector observations with possibly different dimension of their range space are available. At each time any number of these observations (or sensors) can be used in the signal processing performed by the nonlinear filter. The problem considered is the optimal selection of a schedule of these sensors from the available set, so as to optimally estimate a function of the state at the final time. Optimality is measured by a combined performance measure that allocates penalties for errors in estimation, for switching between sensor schedules, and for running a sensor. The solution is obtained in the form of a system of quasi-variational inequalities in the space of solutions of certain Zakai equations.

Key words. nonlinear filtering, sensor scheduling, quasi-variational inequalities

AMS(MOS) subject classifications. 35, 49, 60

1. Introduction.

1.1. Motivation and preliminaries. The problem of nonlinear filtering of diffusion processes has received considerable attention in recent years; see the anthologies [1]-[3] for a review of important developments. In current studies, as well as in related analyses of the partially observed stochastic control problem with such models [4], [5], a key role is played by the linear stochastic partial differential equation describing the evolution of the unnormalized conditional probability measure of the state process given the past of the observations, the so-called Zakai equation.

A significant byproduct of these advances is the feasibility of analyzing complex signal processing problems, including adaptive and sensitivity studies, in an integrated, systematic manner, without heuristic or ad hoc assumptions. A problem of interest in this area is the so-called *sensor scheduling problem.* Roughly speaking, this problem is concerned with the simultaneous selection (according to some performance measure) of a signal processing scheme *together* with the sensors that collect the data to be processed. Particular applications include multiple sensor platforms, distributed sensor networks, and large-scale systems. For example, in a multiple sensor platform, there is definite need for coordinating the data obtained from the various sensors, which may include radar, infrared, or sonar. The data obtained from different sensors are of varying quality and a systematic way is needed for allocating confidence or basing decisions on data collected from different types of sensors. For example, radar sensors are more accurate than infrared sensors for long-range tracking while the opposite is true for short-range tracking. In sensor networks we need to coordinate data collected from a large number of sensors distributed over a large geographical area. Conflicts should be resolved and a preferred set of sensors must be selected over finite (short) time intervals, and used in detection, estimation, or control decisions. Similarly, large-scale systems typically involve an attached information network with the objective of collecting data, processing it, and making the results available to the many control

* Received by the editors January 4, 1988; accepted for publication (in revised form) October 13, 1988.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by United States Army contract DAAG-39-83-C-0028 and by National Science Foundation grants CDR-85-00108 and INT-8413793.

‡ Université Paris Dauphine and Institut National de Recherche en Informatique et en Automatique, 75016 Paris, France.

agents for their decisions (actions). Again the need for coordinating this information in a systematic way is critical.

In such sensor scheduling problems, the systematic utilization of sensors should be the result of optimizing reasonably defined performance measures. Clearly these performance measures will include terms allocating penalties for errors in detection and/or estimation. But more importantly, they must include terms for costs associated with turning sensors on or off, and for switching from one sensor to another. Examples of such costs arising in practice abound. Turning on a radar sensor increases the detectability of the platform (since radars are active sensors) and this should be reflected as a switching cost. Deciding to use a more accurate, albeit more complex, sensor will require higher bandwidth communications and often more computational power allocated to that sensor. In distributed sensor networks it may mean the physical movement of a sensor carrying platform (such as a helicopter or airplane) to a particular geographical location. In large-scale systems the use of several sensors (often hundreds) for decision making may provide better average performance, but it certainly reduces the response speed of the system to changing conditions and increases computational and communication costs both in terms of hardware and software. The latter are obviously evident in large computer/communication networks. These running and switching costs will depend often on the part of the state space occupied by the state vector, i.e., they will be functions of the state as well. For example, sensors have different accuracy or noise characteristics when the state process takes values in different areas of the state space. There is additional cost associated with handling the transfer of information, or tracking record, when there are changes in the set of sensors used; these costs often depend on the state process.

It is not our intent to provide an extensive description of applications here. Detailed descriptions of some of these problems can be found elsewhere; see for example [6], [7]. The underlying thread in all these problem areas is the existence of a variety of sensors, which provide data (for processing), including information of widely varying quality about parameters or variables of interest, for control, detection, estimation, etc. Due to the complexity of these problems it is important to develop systematic conceptual, analytical, and numerical methods for their study and to reduce reliance on ad hoc, heuristic methods as much as possible. The present paper is offered as a contribution in this direction. It provides a general methodology to this problem by reducing it to the analysis of a system of quasi-variational inequalities (see § 3 for details). Numerical methods will be described elsewhere [13].

The sensor scheduling problem is considered here in the context of nonlinear filtering of diffusion processes, and is therefore applicable to detection problems with the same signal models. Modifications of the results apply to other situations including control. In the next section we present a somewhat heuristic definition of the problem, intended to describe the problem clearly, at an intuitive level. The intricacies of establishing this model in a rigorous mathematical fashion are given in § 2 and constitute one of the main contributions of the paper.

1.2. Preliminary description of the problem. The problem considered is as follows. A signal (or state) process $x(\cdot)$ is given, modeled by the diffusion

$$\begin{aligned} dx(t) &= f(x(t))\,dt + g(x(t))\,dw(t), \\ x(0) &= \xi \end{aligned} \tag{1.1}$$

in $\mathbb{R}^n$. We further consider M noisy observations of $x(\cdot)$, described by

$$\begin{aligned} dy^i(t) &= h^i(x(t))\,dt + R_i^{1/2}\,dv^i(t), \\ y^i(0) &= 0 \end{aligned} \tag{1.2}$$

with values in $\mathbb{R}^{d_i}$. Here $w(\cdot)$, $v^i(\cdot)$ are independent, standard, Wiener processes in $\mathbb{R}^n$, $\mathbb{R}^{d_i}$, respectively, and $R_i = R_i^T > 0$ are $d_i \times d_i$ matrices. Further mathematical details on the system (1.1), (1.2) will be given in § 2. Let us consider a finite time horizon $[O, T]$. To formulate the problem of determining an *optimal utilization schedule* for the available sensors, so as to *simultaneously minimize* the cost of errors in estimating a function of $x(\cdot)$ and the costs of using as well as of switching between various sensors, we need to specify these costs. To this end, let $c_i(x)$ denote the cost per unit time when using sensor i, and the state of the system is x; $k_{io}(x)$, $k_{oi}(x)$ denote the cost for turning off, respectively on, the ith sensor when the state of the system is x. The objective of the performed signal processing is to compute, at time T, an estimate $\hat{\phi}(T)$ of a given function $\phi(x(T))$ of the state. Penalties for errors in estimation are assessed according to the cost function

$$E\{c_e(\phi(x(T)) - \hat{\phi}(T))\} := E\{|\phi(x(T)) - \hat{\phi}(T)|^2\}. \tag{1.3}$$

We shall comment briefly on more general estimation problems in § 4 of this paper. In particular, the consideration of a quadratic $c_e(\cdot)$ is not a serious restriction.

Next we consider the set of all possible *sensor activation configurations*, denoted here by $\mathcal{N}$. An element $\nu \in \mathcal{N}$ is a *word* of length M from the alphabet $\{0, 1\}$. If the lth position is occupied by a 1, the lth sensor is activated (used); if by a zero, the lth sensor is off. There are $N = 2^M$ elements in $\mathcal{N}$. A *schedule of sensors is then a piecewise constant function* $u(\cdot) : [O, T] \to \mathcal{N}$. We let $\tau_j \in [O, T]$ denote the instants of changing schedule; i.e., the moments when at least one sensor is turned on or off. At such a switching moment, suppose the schedule before is characterized by $\nu \in \mathcal{N}$, and after by $\nu' \in \mathcal{N}$. Then the *switching cost associated with such a scheduling change* is

$$\mathbf{k}_{\nu\nu'}(x) := \sum_{\{i \in \nu\}\{i \notin \nu'\}} k_{io}(x) + \sum_{\{j \notin \nu\}\{j \in \nu'\}} k_{oj}(x). \tag{1.4}$$

The *total running cost, associated* with schedule $\nu \in \mathcal{N}$ is

$$\mathbf{c}_\nu(x) := \sum_{\{j \in \nu\}} c_j(x). \tag{1.5}$$

In (1.4), (1.5), *the symbol* $\{i \in \nu\}$ *denotes the set of all indices* (*from the set* $\{1, 2, \cdots, M\}$) *that are occupied by a* 1 *in* ν (i.e., the indices corresponding to the sensors that are on); similarly, the symbol $\{i \notin \nu\}$ denotes the set of indices corresponding to sensors that are off.

Using the above notation, the available observations, *under sensor schedule* $u(\cdot)$, are described by

$$dy(t, u(t)) := h(x(t), u(t))\, dt + r(u(t))\, dv(t), \tag{1.6}$$

where it is apparent that the available observations depend explicitly on the sensor schedule $u(\cdot)$. In (1.6), for $x \in \mathbb{R}^n$, $\nu \in \mathcal{N}$,

$$h(x, \nu) := \begin{bmatrix} h^1(x)\chi_{\{\nu\}}(1) \\ \vdots \\ h^i(x)\chi_{\{\nu\}}(i) \\ \vdots \\ h^M(x)\chi_{\{\nu\}}(M) \end{bmatrix}, \tag{1.7}$$

a block column vector, where in standard notation

$$\chi_{\{\nu\}}(i) := \begin{cases} 1 & \text{if the } i\text{th position in the word } \nu \text{ is occupied by a 1} \\ 0 & \text{otherwise.} \end{cases} \tag{1.8}$$

Similarly, for $\nu \in \mathcal{N}$,

$$r(\nu) := \text{block diagonal } \{R_i^{1/2} \chi_{\{\nu\}}(i)\}, \tag{1.9}$$

where R_i are the symmetric, positive matrices defined above. Finally

$$v(t) := \begin{bmatrix} v^1(t) \\ \vdots \\ v^M(t) \end{bmatrix} \tag{1.10}$$

is a higher-dimensional standard Wiener process. In view of (1.7), for all $\nu \in \mathcal{N}$

$$h(\cdot, \nu) : \mathbb{R}^n \to \mathbb{R}^D, \tag{1.11}$$

while

$$r(\nu) : \mathbb{R}^D \to \mathbb{R}^D \tag{1.12}$$

where

$$D = d_1 + d_2 + \cdots + d_M. \tag{1.13}$$

To make the notation clearer, consider the case $M = 2$, $N = 4$. Then $\mathcal{N} = \{00, 01, 10, 11\}$ and

$$\begin{aligned} h(x, 00) &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad h(x, 01) = \begin{bmatrix} 0 \\ h^2(x) \end{bmatrix}, \\ h(x, 10) &= \begin{bmatrix} h^1(x) \\ 0 \end{bmatrix}, \qquad h(x, 11) = \begin{bmatrix} h^1(x) \\ h^2(x) \end{bmatrix}, \end{aligned} \tag{1.14}$$

while

$$\begin{aligned} r(00) &= \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad r(10) = \begin{bmatrix} R_1^{1/2} & 0 \\ 0 & 0 \end{bmatrix}, \\ r(01) &= \begin{bmatrix} 0 & 0 \\ 0 & R_2^{1/2} \end{bmatrix}, \qquad r(11) = \begin{bmatrix} R_1^{1/2} & 0 \\ 0 & R_2^{1/2} \end{bmatrix}. \end{aligned} \tag{1.15}$$

Clearly the dimension of the range space of $y(\cdot, \nu)$ is

$$D_\nu := \sum_{i=1}^{M} d_i \chi_{\{\nu\}}(i). \tag{1.16}$$

Of course for all ν, $y(t, \nu) \in \mathbb{R}^D$.

Following established terminology (cf. [9]), we see that a sensor scheduling strategy is defined by an increasing sequence of switching times $\tau_j \in [O, T]$ and the corresponding sequence $\nu_j \in \mathcal{N}$ of sensor activation configurations. We shall denote such a strategy by $u(\cdot)$, where

$$u(t) = \nu_j, \quad t \in [\tau_j, \tau_{j+1}), \quad j = 1, 2, \cdots. \tag{1.17}$$

As stated earlier we are interested in the *simultaneous* minimization of costs due to estimation errors as well as sensor scheduling. We shall therefore consider *joint estimation and sensor scheduling strategies.* Such a strategy consists of two parts: the

sensor scheduling strategy u (see (1.17)) and the estimator $\hat{\phi}$. The set of admissible strategies U_{ad} is the customary set of strategies adapted to the sequence of σ-algebras

$$\mathcal{F}_t^{y(\cdot,u(\cdot))} := \sigma\{y(s, u(\cdot)), s \leqq t\}. \tag{1.18}$$

That is, we consider *strict sense* admissible controls in the sense of [4]. For the problem under investigation, this last statement must be interpreted very carefully. First, we have indicated in (1.18) that the available past observation data σ-algebra depends (as is evident from (1.6)-(1.9)) very strongly on the sensor schedule $u(\cdot)$. This dependence is nonstandard, as here the dimension of the observation vector and the noise covariance change drastically at each switching time τ_i. In standard stochastic control formulations [4], [5], the dependence of y on $u(\cdot)$ is much more implicit. This is a difficult part of the formulation here, since it prevents us from using Girsanov transformations in a straightforward manner. Second, (1.18) means that the switching times τ_i and the variables ν_i, which define $u(\cdot)$, must be adapted to the filtration $\mathcal{F}_t^{y(\cdot,u(\cdot))}$, *which depends essentially* on the values of τ_i and ν_i! Finally (1.18) also means that $\hat{\phi}(T)$ must be measurable with respect to $\mathcal{F}_t^{y(\cdot,u(\cdot))}$. We describe a rigorous mathematical construction of such a model in § 2.

Given such a strategy, the corresponding cost is

$$J(u(\cdot), \hat{\phi}) := E\Bigg\{ |\phi(x(T)) - \hat{\phi}(T)|^2 \tag{1.19}$$

$$+ \int_0^T c(x(t), u(t))\, dt \tag{1.20}$$

$$+ \sum_j k(x(t), u(\tau_{j-1}), u(\tau_j)) \Bigg\}. \tag{1.21}$$

Here for $x \in \mathbb{R}^n$, $\nu, \nu' \in \mathcal{N}$

$$c(x, \nu) := \mathbf{c}_\nu(x), \tag{1.22}$$

(cf. (1.5)), and

$$k(x, \nu, \nu') = \mathbf{k}_{\nu,\nu'}(x), \tag{1.23}$$

(cf. (1.4)).

The optimal sensor scheduling in nonlinear filtering is thus formulated as the determination of a strategy achieving

$$\inf_{u(\cdot),\hat{\phi}} J(u(\cdot), \hat{\phi}) \tag{1.24}$$

among all admissible strategies.

To somewhat simplify the notation, let us order the elements of $\mathcal{N}$ according to the numbers they represent in binary form. For example in the case $M = 2$, $N = 4$ we replace $\mathcal{N} = \{00, 01, 10, 11\}$ by the set of integers $\{1, 2, 3, 4\}$. That is, the one-to-one correspondence between $\mathcal{N}$ and $\{1, 2, \cdots, N\}$ is described by

$$\begin{aligned} &\nu \mapsto (\text{integer represented by } \nu) + 1, \\ &k \mapsto \text{binary representation of } (k-1). \end{aligned} \tag{1.25}$$

So in the sequel of the paper we replace all the ν, ν' in (1.4)-(1.23) by the corresponding integers from $\{1, 2, \cdots, N\}$.

The structure of the paper is as follows. In § 2 a precise mathematical formulation is given and the corresponding stochastic control problem is precisely defined. In § 3 the set of quasi-variational inequalities solving the problem is derived. In § 4 we offer

some comments and discussion for extensions, further developments, and computational methods.

2. The stochastic control formulation.

2.1. Setting of the model. Let $(\Omega, \mathcal{A}, P)$ be a complete probability space, on which a filtration $\mathcal{F}_t$ is given, $\mathcal{A} = \mathcal{F}_\infty$. Let $w(\cdot)$ and $z(\cdot)$ be two independent, standard $\mathcal{F}_t$-Wiener processes with values in $\mathbb{R}^n$ and $\mathbb{R}^D$, respectively, carried by this probability space. On the same space we consider also an $\mathbb{R}^n$-valued random variable ξ, independent of $w(\cdot)$, $z(\cdot)$, and with probability distribution function π_0.

We consider the Itô equation (1.1), where $f(\cdot)$ is $\mathbb{R}^n$-valued, bounded, and Lipschitz, while $g(\cdot)$ is $\mathbb{R}^{n\times n}$-valued, bounded, and Lipschitz. Letting $a = \frac{1}{2}gg^T$, we assume $a > \alpha I_n$, where $\alpha > 0$ and I_n is the $n \times n$ identity matrix. The Lipschitz property is unnecessary and can be easily removed using Girsanov's transformation (i.e., consider weak solutions of (1.1)) [8]. It is assumed here to simplify the technicalities not related with the main issues of the paper. Under these assumptions (1.1) has a strong solution with well-known properties [8]. Note that *under P, $z(\cdot)$ is independent of $x(\cdot)$.*

Next consider functions $h^i(\cdot)$, $i = 1, \cdots, M$, from $\mathbb{R}^n$ into $\mathbb{R}^{d_i}$ that are bounded and Hölder continuous. We shall denote by L the infinitesimal generator of the Markov process $x(\cdot)$

$$L := \sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2}{\partial x_i\,\partial x_j} + \sum_{i=1}^{n} f_i(x)\frac{\partial}{\partial x_i} \tag{2.1}$$

or in divergence form

$$L := \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x)\frac{\partial}{\partial x_j} - \sum_{i=1}^{n} a_i(x)\frac{\partial}{\partial x_i}, \tag{2.1a}$$

where

$$a_i(x) := -f_i(x) + \sum_{j=1}^{n} \frac{\partial a_{ij}(x)}{\partial x_j}. \tag{2.1b}$$

Let us next consider an *impulsive control* defined as follows. There is a sequence $\tau_1 < \tau_2 \cdots < \tau_k < \cdots$ of increasing $\mathcal{F}_t$-stopping times. To each time τ_i we attach an $\mathcal{F}_{\tau_i}$-measurable random variable u_i with values in the set of integers $\{1, 2, \cdots, N\}$.[1] We define

$$u(t) = u_i, \quad \tau_i \leqq t < \tau_{i+1}, \quad i = 0, 1, 2, \cdots \tag{2.2}$$

and set $\tau_0 = 0$. We require that

$$\tau_i \uparrow T \quad \text{as } i \uparrow \infty, \tag{2.3}$$

while $\tau_k = T$ is possible for some finite k.

Let ν_i be the element of $\mathcal{N}$, corresponding to u_i via (1.25).

Then define

$$h(x, u(t)) := h(x, \nu_i), \qquad \tau_i \leqq t < \tau_{i+1} \tag{2.4}$$

where $h(x, \nu)$ is defined by (1.7), in terms of the given functions $h^i(\cdot)$. Clearly $h(\cdot, u(t))$

[1] Recall that $N = 2^M$ and the binary representation of each integer $1, 2, \cdots, N$ determines a sensor activation configuration by (1.25).

maps $\mathbb{R}^n$ into $\mathbb{R}^D$ for all sensor schedules $u(\cdot)$ and is obviously bounded and Hölder continuous in x. Define also

$$r(u(t)) := r(\nu_i), \qquad \tau_i \leqq t < \tau_{i+1}, \tag{2.5}$$

where $r(\cdot)$ is defined by (1.9), in terms of the given matrices R_i, $i = 1, 2, \cdots, M$. Clearly $r(u(t))$ maps $\mathbb{R}^D$ into $\mathbb{R}^D$ for all sensor schedules $u(\cdot)$ but is *singular*. Next we define $\tilde{h}(x, \nu)$ to be the vector-valued function

$$\tilde{h}(x, \nu) := \begin{bmatrix} R_1^{-1/2} h^1(x) \chi_{\{\nu\}}(1) \\ \vdots \\ R_i^{-1/2} h^i(x) \chi_{\{\nu\}}(i) \\ \vdots \\ R_M^{-1/2} h^M(x) \chi_{\{\nu\}}(M) \end{bmatrix} \tag{2.6}$$

with $\chi_{\{\nu\}}(i)$ defined as in (1.8). Let

$$\tilde{h}(x, u(t)) := \tilde{h}(x, \nu_i), \qquad \tau_i \leqq t < \tau_{i+1}. \tag{2.7}$$

Clearly $\tilde{h}(\cdot, u(t))$ maps $\mathbb{R}^n$ into $\mathbb{R}^D$ for all sensor schedules $u(\cdot)$ and is obviously bounded and Hölder continuous in x. We shall refer to $u(\cdot)$ as the *impulsive control*. As we shall see, it describes essentially the decision to select at a sequence of decision times one of the functions $h(\cdot, k)$, $k \in \{1, 2, \cdots, N\}$. This is the precise mathematical implementation of the sensor selection decision described in the Introduction.

To see that indeed this is the case, we can, with the above preparation, use Girsanov's measure transformation method. Let us then consider the process

$$\zeta(t) = \exp\left\{ \int_0^t \tilde{h}(x(s), u(s))^T \, dz(s) - \frac{1}{2} \int_0^t \|\tilde{h}(x(s), u(s))\|^2 \, ds \right\}, \tag{2.8}$$

where T denotes transpose and $\|\cdot\|$ is the $\mathbb{R}^D$ norm. Note that the process $u(t)$ is adapted to $\mathcal{F}_t$. Then since $x(\cdot)$ is adapted to $\mathcal{F}_t^w \subset \mathcal{F}_t$ and $u(\cdot)$ is cadlag [8], (2.8) is well defined. Moreover, since $\tilde{h}$ is bounded, by Girsanov's theorem [8], [14], $\zeta(\cdot)$ is an $\mathcal{F}_t$-martingale. We can thus define a change of probability measure

$$\left. \frac{dP^{u(\cdot)}}{dP} \right|_{\mathcal{F}_t} = \zeta(t) \tag{2.9}$$

and consider the process

$$v(t) = z(t) - \int_0^t \tilde{h}(x(s), u(s)) \, ds. \tag{2.10}$$

By Girsanov's theorem [8], [14], under the probability measure $P^{u(\cdot)}$ on $(\Omega, \mathcal{A})$, $v(\cdot)$ is a standard $\mathcal{F}_t$-Wiener process with values in $\mathbb{R}^D$. Furthermore, by the independence of $w(\cdot)$ and $z(\cdot)$, $w(\cdot)$ remains a standard $\mathbb{R}^n$-valued, $\mathcal{F}_t$-Wiener process that is independent of $v(\cdot)$. Finally, ξ remains independent of $w(\cdot)$, $v(\cdot)$ while keeping its probability law, denoted by π_0. Thus $x(\cdot)$ also retains its probability law under $P^{u(\cdot)}$.

To relate this construction, i.e., (2.2)–(2.10), to the M noisy observations (sensors) loosely described in the Introduction (cf. in particular (1.6)), observe that (2.10) can be written as

$$r(u(t)) \, dz(t) = h(x(t), u(t)) \, dt + r(u(t)) \, dv(t) \tag{2.11}$$

in view of (1.7), (1.9), (2.4), (2.5), (2.6), and (2.7). Indeed,

$$
\begin{aligned}
r(u(t))\tilde{h}(x,(u(t))) = &\begin{bmatrix} R_1^{1/2}\chi_{\{\nu_i\}}(1) & 0 & 0 \\ 0 & R_2^{1/2}\chi_{\{\nu_i\}}(2) & 0 \\ & & \ddots \\ 0 & 0 & R_M^{1/2}\chi_{\{\nu_i\}}(M) \end{bmatrix} \\
&\cdot \begin{bmatrix} R_1^{-1/2}h^1(x)\chi_{\{\nu_i\}}(1) \\ R_2^{-1/2}h^2(x)\chi_{\{\nu_i\}}(2) \\ \vdots \\ R_M^{-1/2}h^M(x)\chi_{\{\nu_i\}}(M) \end{bmatrix} \\
&= h(x,\nu_i), \qquad \tau_i \leqq t < \tau_{i+1}.
\end{aligned} \tag{2.12}
$$

To give a precise meaning to (1.2), or (1.6), let us introduce the continuous path process in $\mathbb{R}^D$:

$$y(t,u(t)) := y^{\nu_i}(t), \qquad \tau_i \leqq t < \tau_{i+1} \tag{2.13}$$

where

$$dy^{\nu_i}(t) := r(\nu_i)\,dz(t) = h(x(t),\nu_i)\,dt + r(\nu_i)\,dv(t). \tag{2.14}$$

In other words, in the integration from (2.14) to (2.13), we use the left limits of $y(\ ,u(\cdot))$ to initialize. As a consequence, when a sensor is not used, the corresponding components of $y(t,u(t))$ will remain constant, a convention without any consequences. It is clear that if we select $u(t)=\nu$ for all t, where ν has zero everywhere except for one 1 in the ith location, then (1.2) results. It is also rather evident that $dy^{\nu}(t)\in\mathbb{R}^{D_\nu}$ and that in this case the Wiener process $r(\nu)v(\cdot)$ is also D_ν-dimensional (see (1.16) for the definition of D_ν). The process $dy^{\nu_i}(t)$ represents exactly the observation available in $[\tau_i,\tau_{i+1})$.

The next issue we wish to clarify relates to the measurability question we discussed in § 1.2, after (1.18). For any $u(\cdot)$, given the construction of $y(\cdot,u(\cdot))$ above, we can now consider $\mathscr{F}_t^{y(\cdot,u(\cdot))}$ as defined by (1.18). We shall say that $u(\cdot)$ is *admissible*, denoted $u\in U_{\text{ad}}$, if $u(t)$ is $\mathscr{F}_t^{y(\cdot,u(\cdot))}$ measurable, $t>0$, where $\mathscr{F}_t^{y(\cdot,u(\cdot))}$ is constructed as above. More precisely, this means that the τ_i are $\mathscr{F}_t^{u(\cdot,u(\cdot))}$-stopping times or that

$$\{\tau_i < t\} \subset \mathscr{F}_t^{y(\cdot,u(\cdot))} \tag{2.15}$$

and that

$$\nu_i \in \mathscr{F}_{\tau_i}^{y(\cdot,u(\cdot))}. \tag{2.16}$$

Note that since $\mathscr{F}_t^{y(\cdot,u(\cdot))}\subset\mathscr{F}_t$ for any sensor schedule $u(\cdot)$ adapted to $\mathscr{F}_t^{y(\cdot,u(\cdot))}$, if τ_i are $\mathscr{F}_t^{y(\cdot,u(\cdot))}$-stopping times they are also $\mathscr{F}_t$-stopping times, and the above construction (2.8)–(2.14) is still valid. The implication of (2.15), (2.16) is that we should check *that an optimizing strategy, obtained by some procedure, must satisfy the admissibility conditions.* Clearly U_{ad} is nonempty as strategies $u(t)=\nu$, $t\in[O,T]$, obviously are admissible. Also strategies with fixed switchings are admissible. Note that for an admissible control $\mathscr{F}_t^{y(\cdot,u(\cdot))}\subset\mathscr{F}_t^z$. This can be shown in a straightforward manner by proving by induction that $\mathscr{F}_{\tau_i\vee(t\wedge\tau_{i+1})}^{y(\cdot,u(\cdot))}\subset\mathscr{F}_{\tau_i\vee(t\wedge\tau_{i+1})}^z$ using (2.14) and the convention employed in constructing (2.13) from (2.14).

We have thus established in this section the precise mathematical models of nonlinear filtering problems where selection of sensors is possible. In particular we have succeeded in circumventing the subtleties associated with the definition of admissible sensor schedules discussed in § 1.2.[2]

2.2. The optimization problem. For the dynamical system described in § 2.1, we consider now the cost functional (1.19) where the underlying probability measure is $P^{u(\cdot)}$. As indicated in the Introduction, the general problem where the function ϕ will be in a nice class, e.g., bounded C^2, or polynomial, or C^∞ can be treated along identical lines. To simplify the notation we have chosen to formulate the problem for $\phi(x) = x$. The technical difficulties for this case are identical to the ones in the more general cases discussed above, particularly since this $\phi(\cdot)$ is unbounded on $\mathbb{R}^n$. For this choice the selection of the optimal estimator $\hat{\phi}(T)$ is the conditional mean

$$\hat{\phi}(T) = E^{u(\cdot)}\{x(T) \mid \mathcal{F}_T^{y(\cdot, u(\cdot))}\} \tag{2.17}$$

where $E^{u(\cdot)}$ denotes expectation with respect to $P^{u(\cdot)}$. Let $\mu(u, t)$ denote the conditional probability measure of $x(t)$, given $\mathcal{F}_t^{y(\cdot, u(\cdot))}$, on $\mathbb{R}^n$. It is convenient to express (2.17) as a vector valued functional of $\mu(u, t)$:

$$\hat{\phi}(T) = \Phi(\mu(u, T)) = \int_{\mathbb{R}^n} x \, d\mu(u, T). \tag{2.18}$$

We shall further assume that the running and switching cost functions $\mathbf{c}_i(\cdot)$, $\mathbf{k}_{ij}(\cdot)$, $i, j \in \{1, \cdots, N\}$, introduced in (1.4) and (1.5) have the following regularity:

$$\mathbf{c}_i(\cdot), \mathbf{k}_{ij}(\cdot) \quad \text{are in } C_b(\mathbb{R}^n) \text{ (i.e., bounded and continuous).} \tag{2.19}$$

As a result of this simple transformation we can rewrite the cost as a function of the impulsive control $u(\cdot)$ only (i.e., the selection of $\hat{\phi}(\cdot)$ has been eliminated):

$$\begin{aligned} J(u(\cdot)) = E^{u(\cdot)}\Big\{ &\|x(T) - \Phi(\mu(u, T))\|^2 + \int_0^T c(x(t), u(t))\, dt \\ &+ \sum_{j=1}^{\infty} k(x(\tau_j), u(\tau_{j-1}), u(\tau_j)) \chi_{\tau_j < T} \Big\} \end{aligned} \tag{2.20}$$

where $\chi_{\tau_i < T}$ is the characteristic function of the Ω-set $\{\omega;\ \tau_i(\omega) < T\}$. We further assume that the switching costs are uniformly bounded below

$$k(x, i, j) \geqq k_o, \quad x \in \mathbb{R}^n, \quad i, j \in \{1, \cdots, N\} \tag{2.21}$$

with k_0 a positive constant. Note that as a consequence of (2.20) if for some admissible $u(\cdot)$ with positive probability, the number of times $\tau_i < T$ is infinite, then the cost $J(u(\cdot))$ will be infinite. Therefore for T finite the optimal policy will exhibit a finite number of sensor switchings.

The optimal sensor selection problem can now be stated precisely as the optimization problem:

$\mathcal{P}$: Find an admissible impulsive control $u^*(\cdot)$ such that

$$J(u^*(\cdot)) = \inf_{u(\cdot) \in U_{\text{ad}}} J(u(\cdot)) \tag{2.22}$$

[2] Since $r(u(t))$ is a singular matrix, this stage is more delicate than in standard stochastic control theory, where $\mathcal{F}_t^z$ would suffice.

where U_{ad} are all impulsive control strategies adapted to $\mathcal{F}^{y(\cdot),u(\cdot))}$, or equivalently, satisfying (2.15), (2.16). Problem $\mathcal{P}$ is a *nonstandard* stochastic control problem of a partially observed diffusion.

2.3. The equivalent fully-observed problem. In this section we transform the problem of § 2.2 into a fully-observed stochastic control problem by introducing appropriate Zakai equations. As is customary in the theory of nonlinear filtering [1]–[4], we introduce the operator

$$p(u(\cdot), t)(\psi) = E\{\zeta(t)\psi(x(t)) \mid \mathcal{F}_t^{y(\cdot),u(\cdot))}\} \tag{2.23}$$

for each impulsive control $u(\cdot)$. The notation is chosen so as to emphasize the dependence on $u(\cdot)$, which is due to the dependence of $\zeta(\cdot)$ on $u(\cdot)$ as introduced in (2.8).[3] The operator (2.23) maps the set of Borel bounded functions on $\mathbb{R}^n$, into the set of real-valued stochastic processes adapted to $\mathcal{F}_t^{y(\cdot),u(\cdot))}$. Note that $p(u(\cdot), t)$ can be viewed as a positive finite measure on $\mathbb{R}^n$. It is the *unnormalized conditional probability measure* of $x(t)$ given $\mathcal{F}_t^{y(\cdot),u(\cdot))}$ [1], [2].

With the help of these measures we can rewrite the various cost terms in (2.20) as follows:

$$\begin{aligned} E^{u(\cdot)}\{\|x(T) - \Phi(\mu(u, T))\|^2\} &= E\{\zeta(T)\|x(T) - \Phi(\mu(u, T))\|^2\} \\ &= E\{p(u(\cdot), T)(\theta)\} \end{aligned} \tag{2.24}$$

where

$$\theta(x) := \left\| x - \frac{p(u(\cdot), T)(\chi)}{p(u(\cdot), T)(\mathbb{1})} \right\|^2, \tag{2.25}$$

with χ representing the function $\chi(x) := x$ and $\mathbb{1}$ the function $\mathbb{1}(x) := 1$, $x \in \mathbb{R}^n$. A straightforward computation implies that

$$E^{u(\cdot)}\{\|x(T) - \Phi(\mu(u, T))\|^2\} = E\{\Psi(p(u(\cdot), T))\} \tag{2.26}$$

where Ψ is the functional on finite measures of $\mathbb{R}^n$ defined by

$$\Psi(\mu) = \mu(\chi^2) - \frac{\|\mu(\chi)\|^2}{\mu(\mathbb{1})} \tag{2.27}$$

where $\chi^2(x) = \|x\|^2$, $x \in \mathbb{R}^n$, and μ is any finite measure on $\mathbb{R}^n$ such that the quantities $\mu(\chi^2)$ and $\mu(\chi)$ make sense.

Next, we have

$$\begin{aligned} E^{u(\cdot)}\left\{\int_0^T c(x(t), u(t))\, dt\right\} &= E\left\{\zeta(T)\int_0^T c(x, (t), u(t))\, dt\right\} \\ &= E\left\{\int_0^T E\{\zeta(T)c(x(t), u(t)) \mid \mathcal{F}_t\}\, dt\right\} \\ &= E\left\{\int_0^T E\{\zeta(T) \mid \mathcal{F}_t\}c(x(t), u(t))\, dt\right\} \\ &= E\left\{\int_0^T \zeta(t)c(x(t), u(t))\, dt\right\} \end{aligned} \tag{2.28}$$

[3] But the expectation is with respect to P and not $P^{u(\cdot)}$.

because $x(t)$, $u(t)$ are measurable with respect to $\mathcal{F}_t$ and $\zeta(\cdot)$ is an $\mathcal{F}_t$-martingale. Now define a map C with values in $C_b(\mathbb{R}^n)$ via

$$C(u_i) := \mathbf{c}_{u_i}(\cdot), \qquad u_i \in \{1, 2, \cdots, N\}. \tag{2.29}$$

Then in view of (2.29), (2.23), we can rewrite (2.28) as

$$\begin{aligned} E^{u(\cdot)}\left\{\int_0^T c(x(t), u(t))\, dt\right\} &= E\left\{\int_0^T E\{\zeta(t)c(x(t), u(t)) \,|\, \mathcal{F}_t^{y(\cdot, u(\cdot))}\}\, dt\right\} \\ &= E\left\{\int_0^T p(u(\cdot), t)(C(u(t)))\, dt\right\}. \end{aligned} \tag{2.30}$$

Finally,

$$\begin{aligned} E^{u(\cdot)}\{k(x(\tau_i), u(\tau_{i-1}), u(\tau_i))\chi_{\tau_i<T}\} &= E\{\zeta(\tau_i)k(x(\tau_i), u(\tau_{i-1}), u(\tau_i))\chi_{\tau_i<T}\} \\ &= E\{E\{\zeta(\tau_i)k(x(\tau_i), u(\tau_{i-1}), u(\tau_i))\chi_{\tau_i<T} \,|\, \mathcal{F}_{\tau_i}^{y(\cdot, u(\cdot))}\}\} \\ &= E\{p(u(\cdot), \tau_i)(K(u(\tau_{i-1}), u(\tau_i)))\chi_{\tau_i<T}\}. \end{aligned} \tag{2.31}$$

Here we have introduced the function K with values in $C_b(\mathbb{R}^n)$ via

$$K(u_i, u_j) = \mathbf{k}_{u_i, u_j}(\cdot), \qquad u_u, u_j \in \{1, 2, \cdots, N\}, \tag{2.32}$$

and we have used the admissibility of $u(\cdot)$. Note that in the simpler case, where $\mathbf{c}_i(\cdot), \mathbf{k}_{ij}(\cdot)$, $i, j \in \{1, 2, \cdots, N\}$ are constant independent of x, (2.30) simplifies to

$$E^{u(\cdot)}\left\{\int_0^T c(x(t), u(t))\, dt\right\} = E\left\{\int_0^T p(u(\cdot), t)(\mathbb{1})\mathbf{c}_{u(t)}\, dt\right\} \tag{2.33}$$

and (2.31) simplifies to

$$E^{u(\cdot)}\{k(x(\tau_i), u(\tau_{i-1}), u(\tau_i))\chi_{\tau_i<T}\} = E\{\mathbf{k}_{u_{i-1}, u_i}\chi_{\tau_i<T}p(u(\cdot), \tau_i)(\mathbb{1})\}. \tag{2.34}$$

Utilizing (2.26), (2.30), (2.31), we can rewrite the cost corresponding to policy $u(\cdot)$, given in (2.20), as follows:

$$\begin{aligned} J(u(\cdot)) = E\Bigg\{\Psi(p(u(\cdot), T)) &+ \int_0^T p(u(\cdot), t)(C(u(t)))\, dt \\ &+ \sum_{i=1}^{\infty} p(u(\cdot), \tau_i)(K(u_{i-1}, u_i))\chi_{\tau_i<T}\Bigg\}. \end{aligned} \tag{2.35}$$

In (2.35) we have succeeded in displaying the cost as a functional of the unnormalized conditional measure $p(u(\cdot), \cdot)$, which is the "information" state of the equivalent fully-observed stochastic control problem. To complete this transformation we need to derive the evolution equation for $p(u(\cdot), \cdot)$, i.e., the Zakai equation. We turn this problem next and derive a weak form of the Zakai equation for $p(u(\cdot), \cdot)$ in the following lemma. Here $C_b^{2,1}$ denotes the space of all functions $\psi(x, t)$ on $\mathbb{R}^n \times \mathbb{R}$ that are bounded, continuous together with their first and second derivatives with respect to x, and first derivatives with respect to t.

LEMMA 2.1. *For any $\psi \in C_b^{2,1}$ we have the relation*

$$\begin{aligned} p(u(\cdot), t)(\tilde{\psi}(t)) = \pi_0(\tilde{\psi}(0)) &+ \int_0^t p(u(\cdot), s)\left(\frac{\partial \tilde{\psi}}{\partial s} + L\tilde{\psi}\right) ds \\ &+ \int_0^t \sum_{i=1}^{D} p(u(\cdot), s)(\tilde{H}_i(u(s))\tilde{\psi}(s))\, dz_i(s) \end{aligned} \tag{2.36}$$

where

$$
\begin{aligned}
&[\tilde{H}_i(u(s))\phi](x) := \tilde{h}_i(x, u(s))\phi(x), \qquad i = 1, 2, \cdots, D, \quad \phi \in C_b^2, \\
&\tilde{\psi}(s)(x) := \psi(x, s),
\end{aligned} \tag{2.37}
$$

and $\tilde{h}_i$ *is the ith component of* $\tilde{h}$ *(see* (2.6)).

Proof. Let $\beta(\cdot) \in L^\infty(0, T; \mathbb{R}^D)$ given and consider the $\mathcal{F}_t$-martingale $\rho(t)$, defined by

$$
d\rho(t) = \rho(t)\beta(t)^T dz(t), \qquad \rho(0) = 1. \tag{2.38}
$$

Recall that by definition of $\zeta(t)$ (cf. eq. (2.8))

$$
d\zeta(t) = \zeta(t)\tilde{h}(z(t), u(t))^T dz(t), \qquad \zeta(0) = 1. \tag{2.39}
$$

Therefore by Itô's rule [8]

$$
\begin{aligned}
&d(\zeta(t)\rho(t)) = \zeta(t)\rho(t)[(\tilde{h}(x(t), u(t)) + \beta(t))^T dz(t) + \tilde{h}^T(x(t), u(t))\beta(t)\, dt] \\
&\zeta(0)\rho(0) = 1.
\end{aligned} \tag{2.40}
$$

and since $\psi \in C_b^{2,1}$

$$
d\psi(x(t), t) = \left(\frac{\partial \psi(x(t), t)}{\partial t} + L\psi(x(t), t)\right) dt + [\nabla \psi(x(t), t)]^T g(x(t))\, dw(t) \tag{2.41}
$$

where L is given in (2.1). Therefore, with some arguments suppressed for ease of notation

(2.42)

$$
d[\psi(x(t), t)\zeta(t)\rho(t)] = \zeta(t)\rho(t)\left[\left(\frac{\partial \psi}{\partial t} + L\psi + \tilde{h}^T\beta\psi\right) dt + \nabla\psi^T g\, dw(t) + \psi(\tilde{h} + \beta)^T dz(t)\right].
$$

In (2.41), (2.42) we use the notation $\nabla\psi = (\partial\psi/\partial x_1, \cdots, \partial\psi/\partial x_n)^T$. Integrating (2.42), and taking expectations, we deduce

$$
E\{\psi(x(t), t)\zeta(t)\rho(t)\} = \pi_0(\tilde{\psi}(0)) + E\left\{\int_0^t \zeta(s)\rho(s)\left[\frac{\partial \psi}{\partial s} + L\psi + \tilde{h}^T\beta\psi\right] ds\right\}. \tag{2.43}
$$

We can then write

$$
\begin{aligned}
E\left\{\int_0^t \zeta(s)\rho(s)\left[\frac{\partial \psi}{\partial s} + L\psi\right] ds\right\} &= E\left\{\int_0^t E\left\{\rho(s)\zeta(s)\left(\frac{\partial \psi}{\partial s} + L\psi\right) \,\middle|\, \mathcal{F}_s^{y(\cdot, u(\cdot))}\right\} ds\right\} \\
&= E\left\{\int_0^t \rho(s)p(u(\cdot), s)\left(\frac{\partial \psi}{\partial s} + L\psi\right) ds\right\} \\
&= E\left\{\rho(t)\int_0^t p(u(\cdot), s)\left(\frac{\partial \psi}{\partial s} + L\psi\right) ds\right\}
\end{aligned} \tag{2.44}
$$

by virtue of the $\mathcal{F}_t$-martingale property of $\rho(\cdot)$. Similarly,

$$
\begin{aligned}
&E\left\{\int_0^t \zeta(s)\rho(s)\tilde{h}(x(s), u(s))^T\beta(s)\psi(x(s), s)\, ds\right\} \\
&\quad = E\left\{\rho(t)\int_0^t \zeta(s)\psi(x(s), s)\tilde{h}(x(s), u(s))^T dz(s)\right\} \\
&\quad = E\left\{\rho(t)\int_0^t \sum_{i=1}^D p(u(\cdot), s)(\tilde{h}_i(\cdot, u(s))\psi(\cdot, s))\, dz_i(s)\right\}
\end{aligned} \tag{2.45}
$$

where in the first equality we have used the representation $\rho(t) = 1 + \int_0^t \rho(s)\beta(s)^T dz(s)$, and the well-known isomorphism between Itô stochastic integrals and L^2 [8]. Finally,

$$E\{\psi(x(t), t)\zeta(t)\rho(t)\} = E\{\rho(t)p(u(\cdot), t)(\tilde{\psi}(t))\}. \tag{2.46}$$

Using (2.44), (2.45), (2.46) in (2.43), we obtain

$$E\left\{\rho(t)\left[p(u(\cdot), t)(\tilde{\psi}(t)) - \pi_0(\tilde{\psi}(0)) - \int_0^t p(u(\cdot), s)\left(\frac{\partial \psi}{\partial s} + L\psi\right) ds \right.\right. \\ \left.\left. - \int_0^t \sum_{i=1}^{D} p(u(\cdot), s)(\tilde{H}_i(u(s))\tilde{\psi}(s))\, dz_i(s)\right]\right\} = 0. \tag{2.47}$$

We can replace $\rho(t)$ in (2.47) by a linear combination of such variables, with different β. The set of corresponding variables is dense in $L^2(\Omega, \mathcal{F}_t^z, P)$. However, the random variable in the brackets in the right-hand side of (2.47) is clearly in $L^2(\Omega, \mathcal{F}_t^{y(\cdot, u(\cdot))}, P)$ and therefore in $L^2(\Omega, \mathcal{F}_t^z, P)$, since $\mathcal{F}_t^{y(\cdot, u(\cdot))} \subset \mathcal{F}_t^z$. Then (2.47) implies the result of the lemma (2.36).

Remark. Note that the assumed nondegeneracy of $x(\cdot)$ implies that the solution of (2.36) is unique. In general this can be proved under our working hypotheses for solutions that are measure-valued processes. Here we outline such a proof for the case when these conditional measures are absolutely continuous with respect to Lebesgue measure on $\mathbb{R}^n$, i.e., in the case unnormalized conditional densities exist. For this we need to assume in addition that

$$\pi_0 \text{ has a density } p_0 \text{ with respect to Lebesgue measure; } p_0 \in L^2(\mathbb{R}^n). \tag{2.48}$$

We denote by L^* the formal adjoint of L (see (2.1), (2.1a), (2.1b)):

$$L^* = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(x) \frac{\partial}{\partial x_j} + \sum_{i=1}^{n} \frac{\partial}{\partial x_i} a_i, \tag{2.49}$$

and consider the Hilbert space form of the Zakai equation [10]

$$\begin{aligned} dp &= L^* p\, dt + p\tilde{h}(\cdot, u(t))^T dz(t), \\ p(0) &= p_0. \end{aligned} \tag{2.50}$$

The function space in which the solution is sought is

$$L^2(\Omega, \mathcal{A}, P; C(0, T; L^2(\mathbb{R}^n))) \cap L^2_{\mathcal{F}^{y(\cdot, u(\cdot))}}(0, T; H^1(\mathbb{R}^n)). \tag{2.51}$$

Here H^1 is the usual Sobolev space on $\mathbb{R}^n$ [11], and the subindex $\mathcal{F}^{y(\cdot, u(\cdot))}$ in the second L^2 space denotes that the solution is adapted to the filtration $\mathcal{F}_t^{y(\cdot, u(\cdot))}$, $t \geqq 0$. It follows from the results of Pardoux [11] that a unique solution of (2.49) exists in the function space (2.50) under the assumptions made here. We can then establish the following.

LEMMA 2.2. *The following property holds*:

$$p(u(\cdot), t)(\psi) = (p(u(\cdot), t), \psi) \tag{2.52}$$

for all ψ in $L^2(\mathbb{R}^n)$ and bounded, where $(\cdot, \cdot)$ denotes inner product in $L^2(\mathbb{R}^n)$.

Proof. By slight abuse of notation we use the same symbol to denote the conditional unnormalized measure *and* density (whenever the latter exists). Let us prove inductively

that

$$p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1}))(\psi) = (p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1})), \psi), \tag{2.53}$$

where the left-hand notation refers to the measure appearing in (2.36) and the right-hand notation to the solution of (2.50), which is uniquely defined. The induction is necessary because the right-hand side of (2.55) is discontinuous and so we can only examine (2.55) on the intervals $(\tau_i, \tau_i \vee (t \wedge \tau_{i+1}))$. Suppose then that (2.53) holds for $i-1$, and therefore, in particular,

$$p(u(\cdot), \tau_i)(\psi) = (p(u(\cdot), \tau_i), \psi) \quad \forall \psi. \tag{2.54}$$

Now consider the solution η of

$$\begin{aligned} &\frac{\partial \eta}{\partial s} + L\eta = -\eta \tilde{h}(\cdot, u(s))^T \beta(s), \qquad s \in (\tau_i, \tau_i \vee (t \wedge \tau_{i+1})), \\ &\eta(x, \tau_i \vee (t \wedge \tau_{i+1})) = \psi(x) \end{aligned} \tag{2.55}$$

where $\psi \in C_0^\infty(\mathbb{R}^n)$ and β is a smooth deterministic function with values in $\mathbb{R}^D$. From the assumptions on f, g and h^i (it is here that we use the assumed Hölder continuity of h^i), we can assert that the solution of (2.55) belongs to $C_b^{2,1}(\mathbb{R}^n \times (\tau_i, \tau_i \vee (t \wedge \tau_{i+1})))$, for any sample ω [11]. Therefore (2.36) implies (using (2.55))

$$\begin{aligned} p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1}))(\psi) &= p(u(\cdot), \tau_i)(\tilde{\eta}(\tau_i)) \\ &\quad - \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \sum_{j=1}^{D} p(u(\cdot), s)(\tilde{H}_j(u(s))\tilde{\eta}(s))\beta_j(s)\, ds \\ &\quad + \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \sum_{j=1}^{D} p(u(\cdot), s)(\tilde{H}_j(u(s))\tilde{\eta}(s))\, dz_j(s) \end{aligned} \tag{2.56}$$

where $\tilde{H}_j$ is as defined in Lemma 2.1, and $\tilde{\eta}(s)(x) := \eta(x, s)$. Therefore, by Itô's rule and recalling that $\rho(t)$ is the martingale associated with $\beta(t)$, we have

$$\begin{aligned} &p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1}))(\psi)\rho(\tau_i \vee (t \wedge \tau_{i+1})) \\ &\quad = p(u(\cdot), \tau_i)(\tilde{\eta}(\tau_i))\rho(\tau_i) + \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \rho(s) \sum_{j=1}^{D} p(u(\cdot), s)(\tilde{H}_j(u(s))\tilde{\eta}(s))\, dz_j(s) \\ &\qquad + \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \rho(s) \sum_{j=1}^{D} p(u(\cdot), s)(\tilde{H}_j(u(s))\tilde{\eta}(s))\beta_j(s)\, dz_j(s). \end{aligned} \tag{2.57}$$

Hence

$$E\{p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1}))(\psi)\rho(\tau_i \vee (t \wedge \tau_{i+1}))\} = E\{p(u(\cdot), \tau_i)(\tilde{\eta}(\tau_i))\rho(\tau_i)\}. \tag{2.58}$$

On the other hand, from (2.50) and (2.55) we obtain

$$\begin{aligned} (p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1})), \psi) &= (p(u(\cdot), \tau_i), \tilde{\eta}(\tau_i)) \\ &\quad + \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \sum_{j=1}^{D} (p(u(\cdot), s)\tilde{h}_j(\cdot, u(s)), \tilde{\eta}(s))\, dz_j(s) \\ &\quad - \int_{\tau_i}^{\tau_i \vee (t \wedge \tau_{i+1})} \sum_{j=1}^{D} (p(u(\cdot), s), \tilde{H}_j(u(s))\tilde{\eta}(s))\beta_j(s)\, ds, \end{aligned} \tag{2.59}$$

and thus also

$$E\{(p(u(\cdot), \tau_i \vee (t \wedge \tau_{i+1})), \psi)\rho(\tau_i \vee (t \wedge \tau_{i+1}))\} = E\{(p(u(\cdot), \tau_i), \tilde{\eta}(\tau_i))\rho(\tau_i)\}. \tag{2.60}$$

But from the inductive hypothesis (2.54), the right-hand sides of (2.58) and (2.60) are equal. Hence the left-hand sides coincide. Varying β, we easily deduce that (2.53) holds, at least for $\psi \in C_0^\infty(\mathbb{R}^n)$, which is sufficient to conclude the proof of the lemma.

With this result we can rewrite the cost (2.35) as follows:

$$\begin{aligned} J(u(\cdot)) = E\Big\{ & \Psi(p(u(\cdot), T)) + \int_0^T (p(u(\cdot), t), C(u(t)))\, dt \\ & + \sum_{i=1}^{\infty} \chi_{\tau_i < T}(p(u(\cdot), \tau_i), K(u_{i-1}, u_i)) \Big\} \end{aligned} \tag{2.61}$$

where (see (2.27))

$$\Psi(p(u(\cdot), T)) = (p(u(\cdot), T), \chi^2) - \frac{\|(p(u(\cdot), T), \chi)\|^2}{(p(u(\cdot), T), \mathbb{1})}. \tag{2.62}$$

Since (2.62) involves unbounded functions we must show that it makes sense.

At this point it is useful to introduce a weighted Hilbert space to express $\Psi(p(u(\cdot), T))$ in a more convenient form. To this end let

$$\mu(x) = 1 + \|x\|^4 \tag{2.63}$$

and $L^2(\mathbb{R}^n; \mu)$ denote the space of functions φ such that $\varphi\mu \in L^2(\mathbb{R}^n)$. Define in a similar way the space $L^1(\mathbb{R}^n; \mu)$. From the discussion of existence and uniqueness of solutions of (2.50) in the functional space (2.51) and if

$$p_0 \in L^2(\mathbb{R}^n; \mu) \cap L^1(\mathbb{R}^n; \mu),$$

it is easy to check that (2.50), under the assumptions made in § 2.1, has a unique solution in the space

$$L^2(\Omega, \mathcal{A}, P; C(0, T; L^2(\mathbb{R}^n; \mu) \cap L^1(\mathbb{R}^n; \mu))) \cap L^2(0, T; H^1(\mathbb{R}^n; \mu)) \tag{2.64}$$

where $H^1(\mathbb{R}^n; \mu)$ is the obvious modification of $H^1(\mathbb{R}^n)$. This justifies that the quantities arising in (2.62) have a meaning.

We note that $J(u(\cdot))$ is indexed implicitly (we do not include this in our notation) by π_0 (or p_0) and $u(0) = j$, $j \in \{1, \cdots, N\}$, which is deterministic since it is $\mathcal{F}_0^z$-measurable, by construction.

We close this section by rewriting the dynamics (2.50), in terms of the originally given observation nonlinearities h^i, and with forcing inputs the processes $y^i(\cdot)$ introduced in (2.13), (2.14). In view of (2.5), (2.6), (2.7), (2.13), (2.14), we have

$$\tilde{h}(\cdot, u(t))^T\, dz(t) = \sum_{j=1}^{M} h^{j^T}(\cdot)\chi_{\{\nu_i\}}(j) R_j^{-1/2}\, dz_j(t), \qquad \tau_i \leqq t < \tau_{i+1}$$

(where we have written $z = [z_1, z_2, \cdots z_M]^T$)

$$\begin{aligned} &= \sum_{j=1}^{M} h^{j^T}(\cdot)\chi_{\{\nu_i\}}(j) R_j^{-1} R_j^{1/2} \chi_{\{\nu_i\}}(j)\, dz_j(t), \qquad \tau_i \leqq t < \tau_{i+1} \\ &= \delta(\cdot, \nu_i)^T\, dy(t, \nu_i), \qquad \tau_i \leqq t < \tau_{i+1} \\ &=: \delta(\cdot, u(t))^T\, dy(t, u(t)) \end{aligned}$$

where

$$\delta(x,\nu)=\begin{bmatrix} R_1^{-1}h^1(x)\chi_{\{\nu\}}(1) \\ \vdots \\ R_i^{-1}h^i(x)\chi_{\{\nu\}}(u) \\ \vdots \\ R_M^{-1}h^M(x)\chi_{\{\nu\}}(M) \end{bmatrix}. \tag{2.65}$$

Therefore the system dynamics (2.50) can be written equivalently:

$$\begin{aligned} &dp(u(\cdot),t)=L^*p(u(\cdot),t)\,dt+p(u(\cdot),t)\delta(\cdot,u(t))^T\,dy(t,u(\cdot)),\\ &p(u(\cdot),0)=p_0, \end{aligned} \tag{2.66}$$

where $y(t,u(t))$ is defined in (2.13), (2.14). This makes precise the construction of a Zakai equation driven by "controlled" observations alluded to in the Introduction. It also now becomes clear that the spaces described by (2.51), (2.64) are the appropriate ones as far as solutions of (2.50) or (2.66) are concerned.

3. The solution of the optimization problem.

3.1. Setting up a system of quasi-variational inequalities. Let us consider the Banach space $H=L^2(\mathbb{R}^n;\mu)\cap L^1(\mathbb{R}^n;\mu)$ and the metric space H^+ of positive elements of H. Let

$$\begin{aligned} &\mathcal{B}:=\text{space of Borel measurable, bounded functions on } H^+,\\ &\mathcal{C}:=\text{space of uniformly continuous, bounded functions on } H^+. \end{aligned} \tag{3.1}$$

Let us now define semigroups $\Phi_j(t)$ on $\mathcal{B}$ or $\mathcal{C}$ as follows. Consider (2.50) with fixed schedule $u(t)=j$, and let p_j denote the corresponding density $p(\cdot,j)$. Then for $j\in\{1,2,\cdots,N\}$

$$dp_j=L^*p_j\,dt+p_j\tilde{h}^{j^T}\,dz(t),\qquad p_j(0)=\pi \tag{3.2}$$

where

$$\tilde{h}^j:=\tilde{h}(\cdot,j). \tag{3.3}$$

We set

$$\Phi_j(t)(F)(\pi)=E\{F(p_{j,\pi}(t))\},\qquad F\in\mathcal{B}\text{ or }\mathcal{C}, \tag{3.4}$$

where $p_{j,\pi}$ indicates the solution of (3.2) with initial value π. It is easy to see that Φ_j is a semigroup since $p_j(t)$ is a Markov process with values in H^+. It is also useful to introduce the subspaces $\mathcal{B}_1$ and $\mathcal{C}_1$ of functions such that

$$\|F\|_1=\sup_{\pi\in H^+}\frac{|F(\pi)|}{1+\|\pi\|_\mu}<\infty \tag{3.5}$$

where $\|\pi\|_\mu=\|\pi\|_{L^1(\mathbb{R}^n;\mu)}$. The spaces $\mathcal{B}_1$ and $\mathcal{C}_1$ are also Banach spaces. They are needed because we shall encounter functionals with linear growth in the cost function (2.61). To simplify the statement and analysis of the quasi-variational inequalities that solve the optimization problem considered here, we give the details for the case $N=2$ only in the sequel. We shall insert remarks to indicate how the results should be modified for the general case. Let us introduce the notation

$$\begin{aligned} &C_i:=C(i,\cdot),\qquad i=1,2,\\ &K_1:=K(1,2),\qquad K_2:=K(2,1). \end{aligned} \tag{3.6}$$

Since C_1, C_2, K_1, K_2 are bounded functions, we can use them to define elements of $\mathscr{C}_1$ via (for example)

$$C_1(\pi) = (C_1, \pi) \tag{3.7}$$

where a slight abuse of notation, in denoting the functional and the function by the same symbol, has been allowed. Similarly the functional on H^+:

$$\Psi(\pi) = (\pi, \chi^2) - \frac{\|(\pi, \chi)\|^2}{(\pi, \mathbb{1})} \tag{3.8}$$

belongs to $\mathscr{C}_1$ since it is positive and

$$\Psi(\pi) \leqq (\pi, \chi^2) \leqq \|\pi\|_\mu. \tag{3.9}$$

Consider now the set of functionals $U_1(\pi, t)$, $U_2(\pi, t)$ such that

$$\begin{aligned}
&U_1, U_2 \in C(0, T; \mathscr{C}_1),\\
&U_1(\cdot, t) \geqq 0, \qquad U_2(\cdot, t) \geqq 0,\\
&U_1(\pi, T) = U_2(\pi, T) = \Psi(\pi),\\
&U_1(\pi, t) \leqq \Phi_1(s-t) U_1(\pi, s) + \int_t^s \Phi_1(\lambda - t) C_1(\pi)\, d\lambda,\\
&U_2(\pi, t) \leqq \Phi_2(s-t) U_2(\pi, s) + \int_t^s \Phi_2(\lambda - t) C_2(\pi)\, d\lambda \quad \forall s \geqq t,\\
&U_1(\pi, t) \leqq K_1(\pi) + U_2(\pi, t),\\
&U_2(\pi, t) \leqq K_2(\pi) + U_1(\pi, t).
\end{aligned} \tag{3.10}$$

In what follows we occasionally use the notation $U_i(s)(\pi) = U_i(\pi, s)$, $i = 1, 2$.

3.2. Existence of a maximum element. We shall refer to (3.10) as the system of quasi-variational inequalities (QVI). Our first objective is to prove the following.

THEOREM 3.1. *We assume that the conditions on the data f, g, h^i introduced in* § 2.1 *hold. Then the set of functionals U_1, U_2 satisfying* (3.10) *is nonempty and has a maximum element, in the sense that if $\tilde{U}_1, \tilde{U}_2$ denotes this maximum element and U_1, U_2 satisfies* (3.10), *then*

$$\tilde{U}_1 \geqq U_1, \qquad \tilde{U}_2 \geqq U_2.$$

The proof will be carried out in several steps. In fact there is some difficulty due to the functional $\Psi(\pi)$. We shall modify it to assume that

$$0 \leqq \Psi(\pi) \leqq \bar{\Psi}(\pi, \mathbb{1}) \tag{3.11}$$

where $\bar{\Psi}$ is a constant. We shall prove the theorem with the additional assumption (3.11) and prove the probabilistic interpretation, i.e., the connection with the infimum of (2.61). The probabilistic formula will be used next in an approximation procedure. We can approximate, for instance, the functional Ψ defined by (3.8) in the following way. Set

$$\Psi_n(\pi) = \int \frac{\pi \|x\|^2}{1 + (\|x\|^2/n)}\, dx - \frac{\left\| \int (\pi x / ((1 + \|x\|^2/n)^{1/2}))\, dx \right\|^2}{\int \pi\, dx}, \tag{3.12}$$

which clearly satisfies (3.11) with $\bar{\Psi} = n$.

Proof of Theorem 3.1 *under assumption* (3.11). The set of functionals satisfying (3.10) is a subset of $\mathscr{B}_1$ or $\mathscr{C}_1$ defined in (3.5). However for this subset the norm (3.5) is unnecessarily restrictive. For those functionals it is sufficient to set

$$\begin{aligned} &\tilde{H} = L^2(\mathbb{R}^n) \cap L^1(\mathbb{R}^n), \\ &\tilde{H}^+ = \text{set of positive elements of } \tilde{H} \end{aligned} \tag{3.13}$$

and to consider $\tilde{\mathscr{B}}_1$, $\tilde{\mathscr{C}}_1$ the space of Borel or continuous functionals on $\tilde{H}^+$ such that

$$\|F\|_1 = \sup_{\pi \in \tilde{H}^+} \frac{|F(\pi)|}{1 + (\pi, \mathbb{1})} < \infty. \tag{3.14}$$

We shall then study the system (3.10) with $\mathscr{C}_1$ replaced by $\tilde{\mathscr{C}}_1$. Let us note that

$$H^+ \subset \tilde{H}^+,$$

and if we consider a functional F in $\tilde{\mathscr{B}}_1$ or $\tilde{\mathscr{C}}_1$, its restriction to H^+ belongs to $\mathscr{B}_1$ or $\mathscr{C}_1$; the injection

$$F \to \text{restriction of } F \text{ to } H^+$$

is continuous from $\tilde{\mathscr{B}}_1$ or $\tilde{\mathscr{C}}_1$ to $\mathscr{B}_1$ or $\mathscr{C}_1$. Therefore replacing $\mathscr{C}_1$ by $\tilde{\mathscr{C}}_1$ in (3.10) gives a stronger result.

In the proof we shall omit the symbol $\tilde{}$ and write $\mathscr{B}_1$, $\mathscr{C}_1$ instead of $\tilde{\mathscr{B}}_1$, $\tilde{\mathscr{C}}_1$, H^+ instead of $\tilde{H}^+$; the norm $\| \ \|_1$ is then given by (3.14).

The proof is then an adaptation of the methods of Bensoussan and Lions [9] to the present case to take into account the fact that we use $\mathscr{C}_1$ instead of $\mathscr{C}$.

First note that

$$\|\Phi_1(t)\|_{\mathscr{L}(\mathscr{C}_1; \mathscr{C}_1)} \leqq 1 \tag{3.15}$$

where $\mathscr{L}(\mathscr{C}_1; \mathscr{C}_1)$ is the space of linear continuous operators from $\mathscr{C}_1$ into itself. Indeed we have

$$\begin{aligned} \frac{|\Phi_1(t)(F)(\pi)|}{1 + (\pi, \mathbb{1})} &= \frac{|E\{F(p_{1,\pi}(t))\}|}{1 + (\pi, \mathbb{1})} \\ &\leqq \|F\|_1 \frac{(1 + E(p_{1,\pi}(t), \mathbb{1}))}{1 + (\pi, \mathbb{1})} \\ &= \|F\|_1 \end{aligned}$$

since from (3.2)

$$E(p_{1,\pi}(t), \mathbb{1}) = (\pi, \mathbb{1}). \tag{3.16}$$

Therefore

$$\|\Phi_1(t)(F)\|_1 \leqq \|F\|_1, \tag{3.17}$$

which implies (3.15).

Note also that a solution of (3.10) will satisfy

$$U_1(\pi, t) \leqq \Phi_1(T - t) U_1(\pi, T) + \int_t^T \Phi_1(\lambda - t) C_1(\pi) \, d\lambda \tag{3.18}$$

and due to positivity, we also have

$$\|U_1(t)\|_1 \leqq \|U_1(T)\|_1 + \|C_1\|_1 (T - t) \leqq \bar{\Psi} + \|C_1\| (T - t) \tag{3.19}$$

where $\|C_1\| = \sup_x C_1(x)$.

As it is customary in the study of QVI, we begin with the corresponding obstacle problem:

$$
\begin{aligned}
&U_1, U_2 \in C(0, T; \mathscr{C}_1),\\
&U_1(\cdot, t) \geqq 0,\ U_2(\cdot, t) \geqq 0,\\
&U_1(\pi, T) = U_2(\pi, T) = \Psi(\pi),\\
&U_1(\pi, t) \leqq \Phi_1(s-t) U_1(\pi, s) + \int_t^s \Phi_1(\lambda - t) C_1(\pi)\, d\lambda \qquad (3.20)\\
&U_2(\pi, t) \leqq \Phi_2(s-t) U_2(\pi, s) + \int_t^s \Phi_2(\lambda - t) C_2(\pi)\, d\lambda \quad \forall s \geqq t,\\
&U_1(\pi, t) \leqq K_1(\pi) + \zeta_2(\pi, t)\\
&U_2(\pi, t) \leqq K_2(\pi) + \zeta_1(\pi, t)
\end{aligned}
$$

where we assume that

$$
\begin{aligned}
&\zeta_1, \zeta_2 \in C(0, T; \mathscr{C}_1),\\
&\zeta_1(\pi, t) \geqq 0, \qquad \zeta_2(\pi, t) \geqq 0, \qquad (3.21)\\
&\zeta_1(\pi, T), \zeta_2(\pi, T) \geqq \Psi(\pi).
\end{aligned}
$$

We then have the following.

PROPOSITION 3.1. *For ζ_1, ζ_2 as in* (3.21), *the set of U_1, U_2 satisfying* (3.20) *is not empty and has a maximum element.*

It is clear that for ζ_1, ζ_2 given, the system of inequalities (3.20) can be decoupled and U_1, U_2 can be considered separately. Let us then omit indices momentarily and consider

$$
\begin{aligned}
&U \in C(0, T; \mathscr{C}_1),\\
&U(\cdot, t) \geqq 0,\\
&U(\pi, T) = \Psi(\pi),\\
&U(\pi, t) \leqq \Phi(s-t) U(\pi, s) + \int_t^s \Phi(\lambda - t) C(\pi)\, d\lambda \quad \forall s \geqq t, \qquad (3.22)\\
&U(\pi, t) \leqq \zeta(t)
\end{aligned}
$$

where ζ stands, for instance, for $K_1(\pi) + \zeta_2(\pi, t)$. To prove Proposition 3.1, it suffices to show that (3.22) has a maximum element. This can be done by the penalty method. So we look for U_ε solving

(3.23)

$$
\begin{aligned}
&U_\varepsilon(t) = \Phi(s-t) U_\varepsilon(s) + \int_t^s \Phi(\lambda - t) \left[C(\pi) - \frac{1}{\varepsilon} (U_\varepsilon(\lambda) - \zeta(\lambda))^+ \right] d\lambda \quad \text{for } t \leqq s \leqq T,\\
&U_\varepsilon(T)(\pi) = \Psi(\pi),\\
&U_\varepsilon \in C(0, T; \mathscr{C}_1),\\
&U_\varepsilon(\cdot, t) \geqq 0.
\end{aligned}
$$

We can then assert the following lemma.

LEMMA 3.1. *There is a unique solution of* (3.23).

Proof. Note that (3.23) is equivalent to

$$U_\varepsilon(t)=\Phi(T-t)U_\varepsilon(T)+\int_t^T \Phi(\lambda-t)\left[C(\pi)-\frac{1}{\varepsilon}(U_\varepsilon(\lambda)-\zeta(\lambda))^+\right]d\lambda \tag{3.24}$$

and also to

$$\begin{aligned} U_\varepsilon(t) &= e^{-1/\varepsilon(T-t)}\Phi(T-t)\Psi(\pi)+\int_t^T e^{-1/\varepsilon(\lambda-t)}\Phi(\lambda-t) \\ &\quad\cdot\left[C(\pi)+\frac{1}{\varepsilon}U_\varepsilon(\lambda)-\frac{1}{\varepsilon}(U_\varepsilon(\lambda)-\zeta(\lambda))^+\right]d\lambda. \end{aligned} \tag{3.25}$$

Let us define the transformation T_ε of $C(0, T; \mathscr{C}_1)$ into itself using the right-hand side of (3.25). Then the latter can be written as a fixed-point equation:

$$U_\varepsilon = T_\varepsilon U_\varepsilon. \tag{3.26}$$

Using (3.11) and (3.15), we can show precisely, as in Bensoussan and Lions [9, p. 488], that some power of T_ε is a contraction. Hence the result of the lemma follows.

We then can also prove, as in [9, pp. 489–490], that if $\varepsilon \leqq \varepsilon'$, $\|U_\varepsilon\|_1 \leqq K$, then $0 \leqq U_\varepsilon \leqq U_{\varepsilon'}$. As in [9, pp. 494–495] we then show that as $\varepsilon \downarrow 0$, $U_\varepsilon \downarrow U$, which is the maximum element of (3.22). The convergence takes place in $C(0, T; \mathscr{C}_1)$. This establishes Proposition 3.1.

We can then proceed with the proof of Theorem 3.1.

Proof of Theorem 3.1 (*continuation*). Let us consider the map H mapping $C(0, T; \mathscr{C}_1) \times C(0, T; \mathscr{C}_1)$ into itself defined by

$$H(\zeta_1, \zeta_2) = (U_1, U_2) \tag{3.27}$$

where the right-hand side represents the maximum element of (3.20). Now let

$$\begin{aligned} U_1^0(\pi, t) &= \Phi_1(T-t)\Psi(\pi)+\int_t^T \Phi_1(\lambda-t)C_1(\pi)\,d\lambda, \\ U_2^0(\pi, t) &= \Phi_2(T-t)\Psi(\pi)+\int_t^T \Phi_2(\lambda-t)C_2(\pi)\,d\lambda. \end{aligned} \tag{3.28}$$

Consider $\zeta_i(t)$, $\xi_i(t)$, $i = 1, 2$ such that

$$0 \leqq \zeta_i(t) \leqq \xi_i(t) \leqq U_i^o(t), \qquad i = 1, 2, \tag{3.29}$$

$$\xi_i(t) - \zeta_i(t) \leqq \gamma\xi_i(t), \qquad \gamma \in [0, 1]. \tag{3.30}$$

Then we have

$$0 \leqq H(\xi_1, \xi_2) - H(\zeta_1, \zeta_2) \leqq \gamma(1-\gamma')H(\xi_1, \xi_2), \tag{3.31}$$

where

$$\gamma' \leqq \frac{k_0}{k_0 + \bar{\Psi} + \max(\|C_1\|, \|C_2\|)T}. \tag{3.32}$$

Indeed, setting

$$\kappa = 1 - \gamma(1-\gamma'), \tag{3.33}$$

we have to prove that

$$\kappa H(\xi_1, \xi_2) \leqq H(\zeta_1, \zeta_2). \tag{3.34}$$

Let us set

$$(U_1, U_2) = H(\zeta_1, \zeta_2), \qquad (\tilde{U}_1, \tilde{U}_2) = H(\xi_1, \xi_2). \tag{3.35}$$

We need then to show that

$$\kappa\tilde{U}_1 \leqq U_1, \qquad \kappa\tilde{U}_2 \leqq U_2. \tag{3.36}$$

If we can establish that

$$\begin{aligned} \kappa K_1(\pi) + \kappa\xi_2(\pi, t) &\leqq K_1(\pi) + \zeta_2(\pi, t), \\ \kappa K_2(\pi) + \kappa\xi_1(\pi, t) &\leqq K_2(\pi) + \zeta_1(\pi, t), \end{aligned} \tag{3.37}$$

then (3.36) is implied by the monotonicity properties of variational inequalities. But

$$\xi_2(\pi, t)(1-\gamma) \leqq \zeta_2(\pi, t); \tag{3.38}$$

hence, it is enough to establish that

$$\begin{aligned} \kappa K_1(\pi) + \kappa\xi_2(\pi, t) &\leqq K_1(\pi) + (1-\gamma)\xi_2(\pi, t), \\ \kappa K_2(\pi) + \kappa\xi_1(\pi, t) &\leqq K_2(\pi) + (1-\gamma)\xi_1(\pi, t). \end{aligned} \tag{3.39}$$

The first of (3.39) will be satisfied if

$$[\kappa - (1-\gamma)]\xi_2(\pi, t) \leqq (1-\kappa)K_1(\pi) \tag{3.40}$$

or if

$$\gamma'\xi_2(\pi, t) \leqq (1-\gamma')K_1(\pi). \tag{3.41}$$

But observe that

$$\xi_2(\pi, t) \leqq U_2^0(\pi, t) \leqq (\bar{\Psi} + \|C_2\| T)(\pi, \mathbb{1}).$$

So it is enough to choose γ' so that

$$\gamma'(\bar{\Psi} + \|C_2\| T)(\pi, \mathbb{1}) \leqq (1-\gamma')k_0(\pi, \mathbb{1}) \tag{3.42}$$

where k_0 is the uniform lower bound (2.21), since $K_1(\pi) \geqq k_0(\pi, \mathbb{1})$. This last inequality requires

$$\gamma' \leqq \frac{k_0}{k_0 + \bar{\Psi} + \|C_2\| T}. \tag{3.43}$$

In an identical fashion, the second part of (3.39) will be satisfied if

$$\gamma' \leqq \frac{k_0}{k_0 + \bar{\Psi} + \|C_1\| T}. \tag{3.44}$$

So both parts of (3.39) will be satisfied if we choose γ' according to (3.32). The proof of the theorem then proceeds via the standard iteration

$$(U_1^{n+1}, U_2^{n+1}) = H(U_1^n, U_2^n) \tag{3.45}$$

as in [9, pp. 512–514].

Remark. The extension of this result to the general case $N \neq 2$ is straightforward. The system (3.10) has N functionals $U_1, \cdots, U_N$. Everything in (3.10) is the same except for the last two inequalities, which are replaced by

$$U_i(\pi, t) \leqq \min_{\substack{j \neq i \\ j=1,\cdots,N}} (K_{ij}(\pi) + U_j(\pi, t)), \qquad i = 1, \cdots, N. \tag{3.46}$$

We again introduce the system (3.20), where the last two inequalities are replaced by

$$U_i(\pi, t) \leqq \min_{\substack{j \neq i \\ j=1,\cdots,N}} (K_{ij}(\pi) + \zeta_j(\pi, t)), \qquad i = 1, \cdots, N \tag{3.47}$$

where $\zeta_i \in C(0, T; \mathscr{C}_1)$, and satisfy the remainder of (3.21). We then establish the analogue of Proposition 3.1 by penalization. The analogue of Theorem 3.1 is established by introducing a map H mapping $C(0, T; \mathscr{C}_1)^N$ into itself defined by

$$H(\zeta_1, \zeta_2, \cdots, \zeta_N) = (U_1, U_2, \cdots, U_N)$$

where the right-hand side is the maximum element of the analogue of (3.20).

3.3. Existence of an admissible sensor schedule. Our objective in this section is to show that the maximum element U_1, U_2 of the QVI (3.10) provides the value function for the optimization problem (2.61), (2.66) when assumption (3.11) holds. Furthermore, we want to show how an admissible optimal sensor schedule is determined once the pair U_1, U_2 is known.

We shall prove that

$$U_i(\pi, 0) = \inf_{\substack{u(0)=i \\ p(0)=\pi}} J(u(\cdot)), \quad i = 1, 2 \tag{3.48}$$

where $\pi \in H^+$ satisfies $(\pi, \mathbb{1}) = 1$. An optimal schedule will be constructed as follows. To fix ideas, suppose that $i = 1$. Then define

$$\tau_1^* = \inf_{t \leqq T} \{U_1(p_1(t), t) = K_1(p_1(t)) + U_2(p_1(t), t)\} \tag{3.49}$$

where again $p_i(t)$ is the solution of (3.2). We write

$$p^*(t) = p_1(t), \qquad t \in [0, \tau_1^*]. \tag{3.50}$$

Next we define

$$\tau_2^* = \inf_{\tau_1^* \leqq t \leqq T} \{U_2(p_2(t), t) = K_2(p_2(t)) + U_1(p_2(t), t)\}. \tag{3.51}$$

In (3.51), it must be kept in mind that $p_2(t)$ represents the solution of (3.2) with $j = 2$, starting at τ_1^* with value $p_1(\tau_1^*)$. We then define

$$p^*(t) = p_2(t), \qquad t \in [\tau_1^*, \tau_2^*]. \tag{3.52}$$

Note that, unless $\tau_1^* = T$,

$$\tau_2^* > \tau_1^*; \tag{3.53}$$

otherwise

$$\begin{aligned} U_1(p_1(\tau_1^*), \tau_1^*) &= K_1(p_1(\tau_1^*)) + U_2(p_1(\tau_1^*), \tau_1^*), \\ U_2(p_1(\tau_1^*), \tau_1^*) &= K_2(p_1(\tau_1^*)) + U_1(p_1(\tau_1^*), \tau_1^*), \end{aligned} \tag{3.54}$$

which is impossible since

$$K_1(p_1(\tau_1^*)) > 0, \qquad K_2(p_1(\tau_1^*)) > 0 \quad \text{a.s.} \tag{3.55}$$

Similarly we proceed to construct a sequence of $\tau_1^* < \tau_2^* < \tau_3^* < \cdots$ and the process $p^*(\cdot)$. We can then prove the following.

THEOREM 3.2. *With the same assumptions as in Theorem* 3.1, *and in addition, assuming that* (3.11) *holds, the sequence of stopping times* $\tau_1^*, \tau_2^*, \cdots$ *defines an optimal admissible sensor schedule.*

Proof. Considering (3.10) as a VI with obstacle ζ_2, ζ_1, we can write from the definition of τ_1^*:

$$U_1(\pi, 0) = E\left\{ U_1(p_1(\tau_1^*), \tau_1^*) + \int_0^{\tau_1^*} C_1(p_1(\lambda))\, d\lambda \right\}. \tag{3.56}$$

This can be established by using the penalization (3.23), along lines similar to those of [9, pp. 578-587]. Then

$$\begin{aligned} E\{U_1(p_1(\tau_1^*), \tau_1^*)\} &= E\{U_1(p^*(\tau_1^*)), \tau_1^*\} \\ &= E\{\Psi(p^*(T))\chi_{\tau_1^*=T}\} + E\{U_1(p^*(\tau_1^*), \tau_1^*)\chi_{\tau_1^*<T}\}. \end{aligned}$$

Substituting back in (3.56) and using the definition of τ_1^* in (3.49), we obtain

$$\begin{aligned} U_1(\pi, 0) = E\Bigg\{ &\Psi(p^*(T))\chi_{\tau_1^*=T} + \int_0^{\tau_1^*} C_1(p^*(\lambda))\, d\lambda \\ &+ K_1(p^*(\tau_1^*))\chi_{\tau_1^*<T} + U_2(p^*(\tau_1^*), \tau_1^*)\chi_{\tau_1^*<T} \Bigg\}. \end{aligned} \tag{3.57}$$

Furthermore, again by employing penalization, we can show that

$$E\{U_2(p^*(\tau_1), \tau_1^*)\} = E\{U_2(p_2(\tau_1^*), \tau_1^*)\} = E\left\{ U_2(p_2(\tau_2^*), \tau_2^*) + \int_{\tau_1^*}^{\tau_2^*} C_2(p_2(\lambda))\, d\lambda \right\}. \tag{3.58}$$

This implies

$$E\{U_2(p_2(\tau_1^*), \tau_1^*)\chi_{\tau_1^*<T}\} = E\left\{ U_2(p_2(\tau_2^*), \tau_2^*)\chi_{\tau_1^*<T} + \chi_{\tau_1^*<T} \int_{\tau_1^*}^{\tau_2^*} C_2(p_2(\lambda))\, d\lambda \right\}. \tag{3.59}$$

Next

$$E\{U_2(p_2(\tau_2^*), \tau_2^*)\chi_{\tau_1^*<T}\} = E\{\Psi(p^*(T))\chi_{\tau_1^*<T, \tau_2^*=T}\} + E\{U_2(p^*(\tau_2^*), \tau_2^*)\chi_{\tau_2^*<T}\}.$$

Substituting back in (3.57) and using the definition of τ_2^* in (3.51), we obtain

$$\begin{aligned} U_1(\pi, 0) = E\Bigg\{ &\Psi(p^*(T))\chi_{\tau_2^*=T} + K_1(p^*(\tau_1^*))\chi_{\tau_1^*<T} + K_2(p^*(\tau_2^*))\chi_{\tau_2^*<T} \\ &+ \int_0^{\tau_1^*} C_1(p^*(\lambda))\, d\lambda + \int_{\tau_1^*}^{\tau_2^*} C_2(p^*(\lambda))\, d\lambda + U_1(p^*(\tau_2^*), \tau_2^*)\chi_{\tau_2^*<T} \Bigg\}. \end{aligned} \tag{3.60}$$

Proceeding in a similar fashion and collecting results we can write:

$$\begin{aligned} U_1(\pi, 0) = E\Bigg\{ &\Psi(p^*(T))\chi_{\tau_n^*=T} + \sum_{i=1}^{n} K_i(p^*(\tau_i^*))\chi_{\tau_i^*<T} \\ &+ \sum_{i=0}^{n-1} \chi_{\tau_{i+1}^*<T} \int_{\tau_i^*}^{\tau_{i+1}^*} C_{i+1}(p^*(\lambda))\, d\lambda \Bigg\} \\ &+ U_{n+1}(p^*(\tau_n^*), \tau_n^*)\chi_{\tau_n^*<T} \end{aligned} \tag{3.61}$$

where we use the notation

$$\begin{aligned} K_i &= \begin{cases} K_1 & \text{if } i \text{ is odd,} \\ K_2 & \text{if } i \text{ is even,} \end{cases} \\ C_i &= \begin{cases} C_1 & \text{if } i \text{ is odd,} \\ C_2 & \text{if } i \text{ is even,} \end{cases} \\ U_i &= \begin{cases} U_1 & \text{if } i \text{ is odd,} \\ U_2 & \text{if } i \text{ is even.} \end{cases} \end{aligned} \tag{3.62}$$

However, observe that necessarily $\tau_n^* = T$ for n large enough (random). Otherwise we have $\tau_n^* < T$ for all n, on a set $\Omega_0 \subset \Omega$ of positive probability. But $\tau_n^* \uparrow \tau^* \leqq T$ and

$$(p^*(\tau_i^*), \mathbb{1}) \to (p^*(\tau^*), \mathbb{1}) \tag{3.63}$$

where (since $(\pi, \mathbb{1}) = 1$)

$$(p^*(\tau^*), \mathbb{1}) = 1 + \int_0^{\tau^*} p^* \delta^T dy \tag{3.64}$$

(see (2.66)) and

$$(p^*(\tau^*), \mathbb{1}) = E\{\zeta(\tau^*) \mid \mathscr{F}_{\tau^*}^{y(\cdot, u^*)}\} > 0 \quad \text{a.s.} \tag{3.65}$$

where $\zeta(\cdot)$ is the process introduced by (2.8). Therefore on Ω_0, as $n \to \infty$

$$\sum_{i=1}^{n} K_i(p^*(\tau_i^*)) \chi_{\tau_i^* < T} \to +\infty \tag{3.66}$$

and since Ω_0 has positive probability, as $n \to \infty$

$$E\left\{\sum_{i=1}^{n} K_i(p^*(\tau_i^*)) \chi_{\tau_i^* < T}\right\} \to \infty, \tag{3.67}$$

which contradicts (3.19).

We can thus assert that

$$\chi_{\tau_n^* = T} \to 1 \quad \text{a.s.} \tag{3.68}$$

In particular, it follows that the sequence $\tau_1^*, \tau_2^*, \cdots$, defines an admissible schedule denoted by u^*. The corresponding state solution of (2.66) coincides with p^* and (3.61) implies

$$U_1(\pi, 0) \geqq J(u^*(\cdot)). \tag{3.69}$$

But by standard arguments, we check that

$$U_1(\pi, 0) \leqq J(u(\cdot)) \quad \forall u(\cdot) \in U_{\text{ad}} \tag{3.70}$$

and therefore $u^*(\cdot)$ is indeed optimal.

3.4. The main result. We want now to get rid of (3.11) and consider the original functional Ψ in (3.8). Let us consider the approximation (3.12) Ψ_n of Ψ. To Ψ_n corresponds a system of QVI:

$$\begin{aligned}
&U_1^n, U_2^n \in C(0, T; \tilde{\mathscr{C}}_1),\\
&U_1^n, U_2^n \geqq 0,\\
&U_1^n(\pi, T) = U_2^n(\pi, T) = \Psi_n(\pi),\\
&U_1^n(\pi, t) \leqq \Phi_1(s-t) U_1^n(\pi, s) + \int_t^s \Phi_1(\lambda - t) C_1(\pi)\, d\lambda,\\
&U_2^n(\pi, t) \leqq \Phi_2(s-t) U_2^n(\pi, s) + \int_t^s \Phi_2(\lambda - t) C_2(\pi)\, d\lambda \quad \forall s \geqq t\\
&U_1^n(\pi, t) \leqq K_1(\pi) + U_2^n(\pi, t),\\
&U_2^n(\pi, t) \leqq K_2(\pi) + U_1^n(\pi, t).
\end{aligned} \tag{3.71}$$

From Theorem 3.2, we can assert that

$$U_i^n(\pi, 0) = \inf_{\substack{u(0)=i \\ p(0)=\pi}} J^n(u(\cdot)), \qquad i = 1, 2 \tag{3.72}$$

where

$$J^n(u(\cdot)) = E\left\{\Psi^n(p(u(\cdot), T)) + \int_0^T (p(u(\cdot), t), C(u(t)))\, dt + \sum_{i=1}^{\infty} \chi_{\tau_i < T}(p(u(\cdot), \tau_i), K(u_{i-1}, u_i))\right\}. \tag{3.73}$$

Therefore we deduce that

$$J^n(u(\cdot)) - J(u(\cdot)) = E\{\Psi_n(p(u(\cdot), T)) - \Psi(p(u(\cdot), T))\} \tag{3.74}$$

and from (3.12) we deduce

$$\begin{aligned} |J^n(u(\cdot)) - J(u(\cdot))| \leqq{}& E\left\{\int \frac{p(u(\cdot), T)\|x\|^4}{n + \|x\|^2}\, dx\right\} \\ &+ E\left\{\left(\int p(u(\cdot), T)x\left(1 - \frac{1}{(1 + \|x\|^2/n)^{1/2}}\right) dx\right)^T \right. \\ &\cdot \left(\int p(u(\cdot), T)x\left(1 + \frac{1}{(1 + \|x\|^2/n)^{1/2}}\right) dx\right) \\ &\left. \cdot \frac{1}{\int p(u(\cdot), T)\, dx}\right\}. \end{aligned} \tag{3.75}$$

But using (2.50) yields (see (2.1a))

$$\begin{aligned} &E\left\{\int \frac{p(u(\cdot), t)\|x\|^4}{n + \|x\|^2}\, dx\right\} \\ &= E\left\{\int_0^t \int p(u(\cdot), s)(x)\left\{\frac{\partial a_{ij}}{\partial x_i} \frac{2\|x\|^2(2n + \|x\|^2)x_j}{(n + |x|^2)^2}\right.\right. \\ &\qquad + a_{ij}\left(\delta_{ij}\frac{2\|x\|^2(2n + \|x\|^2)}{(n + \|x\|^2)^2} + \frac{8x_i x_j n^2}{(n + \|x\|^2)^3}\right) \\ &\qquad \left.\left. - a_i \frac{2\|x\|^2(2n + \|x\|^2)x_i}{(n + \|x\|^2)^2}\right\} ds + \int \frac{\pi(x)\|x\|^4}{n + \|x\|^2}\, dx\right\} \end{aligned}$$

where we employ the summation convention over repeated indices. Hence after majorizing conveniently, we have

$$\begin{aligned} &E\left\{\int \frac{p(u(\cdot), t)(x)\|x\|^4}{n + \|x\|^2}\, dx\right\} \\ &\leqq \int \frac{\pi(x)\|x\|^4}{n + \|x\|^2}\, dx + \Gamma \int_0^t E\left\{\int \frac{p(u(\cdot), s)(x)\|x\|^4}{n + \|x\|^2}\, dx\right\} ds + \frac{\Gamma t}{n}. \end{aligned} \tag{3.76}$$

We shall use capital Greek letters, $\Gamma, \Delta, \cdots$, to indicate constants in the following estimates. Finally we deduce that

$$E\left\{\int \frac{p(u(\cdot),t)(x)\|x\|^4}{n+\|x\|^2}\,dx\right\} \leqq \Gamma_t\left[\int \frac{\pi(x)\|x\|^4}{n+\|x\|^2}\,dx+\frac{1}{n}\right] \leqq \Gamma_t\left[\frac{1}{n}\int \pi(x)\|x\|^4\,dx+\frac{1}{n}\right]. \tag{3.77}$$

Next consider

$$\frac{p(u(\cdot),t)}{(p(u(\cdot),t),\mathbb{1})} = \sigma(u(\cdot),t),$$

which is the normalized conditional probability measure and satisfies Kushner's equation

$$d(\sigma(t)(\varphi)) = \sigma(t)(L\varphi)\,dt + (\sigma(t)(\tilde{h}\varphi) - \sigma(t)(\varphi)\sigma(t)(\tilde{h}))\cdot(dz - \sigma(t)(\tilde{h})\,dt). \tag{3.78}$$

If we apply (3.78) with $\varphi = \|x\|^2 = \chi^2$, we obtain

$$\begin{aligned} dE\{\sigma(t)(\chi^2)\} &= E\{\sigma(t)(L\chi^2) - \sigma(t)(\tilde{h})[\sigma(t)(\tilde{h}\chi^2) - \sigma(t)(\chi^2)\sigma(t)(\tilde{h})]\}\,dt \\ &\leqq \Delta_0(1+E\{\sigma(t)(\chi^2)\})\,dt. \end{aligned} \tag{3.79}$$

Finally,

$$E\{\sigma(t)(\chi^2)\} \leqq \Delta_t \int \pi(x)\|x\|^2\,dx.$$

But the second term in (3.75) is

$$\begin{aligned} &E\left\{\sigma(T)\left(\chi\left(1+\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)^T (p(T))\left(\chi\left(1-\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right\} \\ &\quad\leqq \left[E\left\{\left\|\sigma(T)\left(\chi\left(1+\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right\|^2\right\}\right]^{1/2} \\ &\qquad\cdot\left[E\left\{\left\|p(T)\left(\chi\left(1-\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right\|^2\right\}\right]^{1/2} \\ &\quad\leqq \Delta^1(E\{\sigma(T)(\chi^2)\})^{1/2}\left(E\left\{\left\|p(T)\left(\chi\left(1-\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right\|^2\right\}\right)^{1/2} \\ &\quad\leqq \Delta^2\left[E\left\{\left\|p(T)\left(\chi\left(1-\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right\|^2\right\}\right]^{1/2} \\ &\quad= \Delta^3\left[E\left\{\sum_i\left(p(T)\left(\chi_i\left(1-\frac{1}{(1+\chi^2/n)^{1/2}}\right)\right)\right)^2\right\}\right]^{1/2} \\ &\quad\leqq \Delta^3\left[E\left\{p(T)(\chi^2)p(T)\left(\frac{\chi^2}{n+\chi^2}\right)\right\}\right]^{1/2}. \end{aligned} \tag{3.81}$$

We easily check that

$$E\{(p(T)(\chi^2))^2\} \leqq \Delta^4 + \left(\int \pi(x)\|x\|^2\,dx\right)^2 \leqq \Delta^5,$$

$$\begin{aligned} dE\left\{\left|p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right|^2\right\} &\leqq 2E\left\{p(t)\left(L\frac{\chi^2}{n+\chi^2}\right)p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right\}dt \\ &\quad+\Delta^5 E\left\{\left|p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right|^2\right\}dt. \end{aligned}$$

But

$$L\frac{\chi^2}{n+\chi^2}\leqq\frac{\Delta^6}{\sqrt{n}}; \tag{3.82}$$

hence

$$dE\left\{\left|p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right|^2\right\}\leqq\left[\Delta^5 E\left\{\left|p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right|^2\right\}+\frac{\Delta^7}{n}\right]dt, \tag{3.83}$$

which implies

$$\begin{aligned} E\left\{\left|p(t)\left(\frac{\chi^2}{n+\chi^2}\right)\right|\right\}&\leqq\Theta_t\left[\frac{1}{n}+\left(\int\frac{\pi(x)\|x\|^2\,dx}{n+\|x\|^2}\right)^2\right]\\ &\leqq\frac{\Theta_t}{n}\left(1+\int\pi(x)\|x\|^4\,dx\right). \end{aligned} \tag{3.84}$$

Therefore, continuing from (3.81), the second term in (3.75) is majorized by $\Gamma_0/n^{1/4}$. Collecting results (from (3.75), (3.77), (3.81), (3.84)), we can assert that

$$|J^n(u(\cdot))-J(u(\cdot))|\leqq\frac{\Delta}{n^{1/4}} \tag{3.85}$$

provided the initial distribution of $p(0)$, i.e., π satisfies

$$\int\pi(x)\|x\|^4\,dx<\infty. \tag{3.86}$$

The estimate in (3.85) is uniform with respect to n. Therefore

$$\left|U_i^n(\pi,0)-\inf_{u(0)=i,p(0)=\pi}J(u(\cdot))\right|\leqq\frac{\Delta}{n^{1/4}}. \tag{3.87}$$

In fact we can replace zero by any $t\in[0,T]$ and consider the function

$$U_i(\pi,t)=\inf_{u(t)=i,p(t)=\pi}J_t(u(\cdot)) \tag{3.88}$$

where $J_t(u(\cdot))$ corresponds to a problem analogous to (2.50), (2.61) starting in t instead of zero. Therefore we have

$$|U_i^n(\pi,t)-U_i(\pi,t)|\leqq\frac{\Delta}{n^{1/4}}. \tag{3.89}$$

However we must be careful of the fact that the constant in (3.89) depends on a bound on $\int\pi(x)\|x\|^4\,dx$. More precisely, we have proved that

$$U_i^n(\pi,t)-U_i(\pi,t)|\leqq\frac{\Delta'}{n^{1/4}}\left(1+\int\pi(x)\|x\|^4\,dx\right) \tag{3.90}$$

where Δ' here does not depend on π (assuming that π is a probability). It follows that

$$U_i^n(\pi,t)\to U_i(\pi,t)\quad\text{in } C(0,T;\mathscr{C}_1). \tag{3.91}$$

Taking the limit in (3.71), we obtain that U_1, U_2 is a solution of (3.10), and moreover

$$U_i(\pi,0)=\inf_{\substack{u(0)=i\\p(0)=\pi}}J(u(\cdot)). \tag{3.92}$$

However, by a probabilistic argument already used in § 3.3, any solution of (3.10) is smaller than the right-hand side of (3.92). This completes the proof of Theorem 3.1, and also provides the same statement as in Theorem 3.2, without assumption (3.11) and for our original Ψ given by (3.8).

REFERENCES

[1] M. HAZEWINKEL AND J. C. WILLEMS, EDS., *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, Proc. NATO Advanced Study Institute, Les Arcs, France, Reidel, Dordrecht, The Netherlands, 1981.

[2] W. H. FLEMING AND L. G. GOROSTIZA, EDS., *Advances in Filtering and Optimal Stochastic Control*, Proc. IFIP-WG 7/1 Working Conference, Cocoyoc, Mexico, 1982, Lecture Notes in Control and Information Sci. 42, Springer-Verlag, Berlin, New York, 1982.

[3] M. METIVIER AND E. PARDOUX, EDS., *Stochastic Differential Systems: Filtering and Control*, Proc. IFIP-WG 7/1 Working Conference, Marseille-Luminy, France, 1984, Lecture Notes in Control and Information Sci. 69, Springer-Verlag, Berlin, New York, 1985.

[4] W. H. FLEMING AND E. PARDOUX, *Optimal control for partially observed diffusions*, SIAM J. Control Optim., 20 (1982), pp. 261–285.

[5] J. M. BISMUT, *Partially observed diffusions and their control*, SIAM J. Control Optim., 20 (1982), pp. 302–309.

[6] J. S. BARAS, *Optimal sensor scheduling in multiple sensor platforms*, in preparation.

[7] J. D. KATTAR, *A solution of the multi-weapon, multi-target assignment problem*, WP-26597, The MITRE Co., Bedford, MA, February 1986.

[8] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, Berlin, New York, 1980.

[9] A. BENSOUSSAN AND J. L. LIONS, *Contrôle impulsionnel et inéquations quasi-variationelles*, Dunod, Paris, 1982.

[10] A. BENSOUSSAN, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169–222.

[11] E. PARDOUX, *Stochastic partial differential equations and filtering of diffusion processes*, Stochastics, 9 (1983).

[12] J. L. LIONS AND E. MAGENES, *Problèmes aux limites non homogènes et applications*, Vols. 1 and 2, Dunod, Paris, 1968; English translation, Springer-Verlag, Berlin, New York, 1972.

[13] J. S. BARAS AND A. BENSOUSSAN, *Optimal sensor scheduling in nonlinear filtering of diffusion processes II: computational methods*, in preparation.

[14] R. S. LIPSTER AND A. N. SHIRYAYEV, *Statistics of Random Processes* I: *General Theory*, Springer-Verlag, Berlin, New York, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 814–835, July 1989

008

FREQUENCY-SCALE DECOMPOSITION OF H^∞-DISK PROBLEMS*

D. WILLIAM LUSE† AND JOSEPH A. BALL‡

Abstract. Time- and frequency-scale decomposition methods have been used extensively for the simplification of automatic control problems. This paper considers the problem of parametrizing all functions in a specified H^∞-function disk over the unit ball in H^∞. This problem arises in the design of stabilizing feedback compensators to minimize a weighted sensitivity matrix in the H^∞-norm sense. It is shown that an H^∞-disk problem, whose data has two-frequency-scale behavior, can be broken down into slow and fast subproblems. If solutions can be found for both subproblems, then the solutions can be combined to give an approximate solution for the original problem.

Key words. singular perturbation, optimal control, factorization theory, Beurling–Lax Theorem

AMS(MOS) subject classifications. 47A68, 93B28, 93B35, 93D15

1. Introduction.

Frequency-domain plots have been used for many years to represent control system specifications and to aid in the design of feedback compensators. The use of such plots has proved very useful for scalar systems. For multivariable systems, frequency-domain plots still have a strong physical interpretation [1]; but compensator design based on them seems to require a great deal of experience on the part of the designer. Thus, there is a pressing need to automate frequency-domain system design. As pointed out in [2], the constraints diagrammed in frequency-domain plots can usually be translated mathematically into "function disk" inclusions, in which the "disks" are described in terms of H^∞ and L^∞ norms. The development of [2] further notes the existence of a well-developed mathematical theory for treating problems of this type. We now consider the problem (e.g., [3]) of choosing a stabilizing compensator that minimizes the weighted sensitivity of a feedback loop with respect to open-loop plant perturbations. The minimization is to be done in the H^∞ sense. The following notation will be used. $\mathbb{C}$ will denote the complex numbers and $\mathbb{C}^+$ will denote the open half-plane $\operatorname{Re}(s) > 0$. $\mathbb{R}$ will denote the real numbers. H^∞ and H^2 are as defined in [4] and [2]. The notations S and $M(\cdot)$ are borrowed from [3]. S is the ring of proper, stable (bounded on C^+), real rational functions of the complex variable s. We can give S a norm by imbedding it in H^∞. In other words, S is "real rational H^∞." $M(\cdot)$ is shorthand notation for the set of all matrices whose elements lie in some set. Thus, $M(S)$ stands for the set of all rational, proper, stable transfer matrices. When the $M(\cdot)$ notation is used, it is assumed that the sizes of the matrices can be determined by context. Borrowing standard notation from algebra, the field of real rational functions of s is $\mathbb{R}(s)$. The sensitivity minimization problem [3] can now be formulated. Referring to Fig. 1.1, we assume that $Q(s) \in M(\mathbb{R}(s))$ is given and that we wish to find $C(s) \in M(\mathbb{R}(s))$ such that the sensitivity $S(s)$ given by (1.1) is a minimum, and the (four-block) closed-loop transfer matrix of Fig. 1.1 is an element of $M(S)$. We assume in this paper that Q is square:

$$S(s) = (I + Q(s)C(s))^{-1}. \tag{1.1}$$

The sense of the minimum is the weighted H^∞ norm sense. Thus, the problem is to

* Received by the editors January 18, 1988; accepted for publication (in revised form) September 13, 1988.

† Bradley Department of Electrical Engineering, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

‡ Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061.

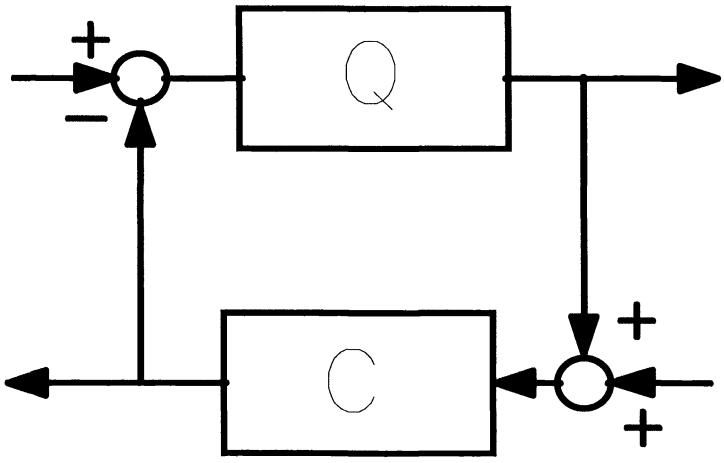

FIG. 1.1

find a compensator $C(s)$ that leaves the closed-loop system of Fig. 1.1 stable and makes the quantity $\rho(C)$ given in (1.2) a minimum:

$$\begin{aligned}\rho(C) &= \sup_{\mathrm{Re}\,(s)>0} \bar{\sigma}(W_1(s)S(s)W_2(s)) \\ &= \sup_{\omega \in R} \bar{\sigma}(W_1(j\omega)S(j\omega)W_2(j\omega)) \\ &= \|W_1 S W_2\|_\infty .\end{aligned} \tag{1.2}$$

We assume that W_1 and W_2 are rational matrix functions such that $W_j \in M(S)$ and $W_j^{-1} \in M(S)$ for $j = 1, 2$.

It is a matter of controversy whether this problem is actually one that control engineers want to solve. In general, the optimal compensator is not physically realizable. Sometimes a sequence of compensators can be found for which the quantity (1.2) converges to the optimal value. This is not true, however, in the case of a strictly proper, stable, scalar plant. In this case, the optimal value of (1.2) is zero, while the minimum realizable value of (1.2) is at least $\|W_1(\infty)W_2(\infty)\|$. This follows because $C(\infty) = 0$ for any realizable compensator. In general, the optimal value of (1.2) can be attained by using a physically realizable compensator only when $Q(s)$ has no poles or zeros on the extended imaginary axis. In [5] it is demonstrated that control engineers often want to make tradeoffs between several quantities similar to (1.2). There are often several function disks within which the control system designer wants various system transfer matrices to lie. Unfortunately, there is no known solution to such "multidisk" problems. Thus, the sensitivity minimization problem described here can be viewed either as an intermediate step toward further research or as a design tool that may aid in evaluating candidate system designs.

The sensitivity minimization problem as stated above has been solved in a number of different ways. In this paper, we use the approach of [6]. The first step toward solution is to rewrite the sensitivity minimization problem as a matrix interpolation problem. Let Q be expressed as left- and right-coprime matrix fractions over $M(S)$ as shown in (1.3) [3]:

$$Q = ND^{-1} = \tilde{D}^{-1}\tilde{N}, \tag{1.3}$$

$$XN + YD = I, \tag{1.4a}$$

$$\tilde{N}\tilde{X} + \tilde{D}\tilde{Y} = I. \tag{1.4b}$$

Let $X, Y, \tilde{X}, \tilde{Y} \in M(S)$ be such that the Bezout equations (1.4) hold. Then any stabilizing compensator can be expressed as

$$C = (Y - R\tilde{N})^{-1}(X + R\tilde{D}) \tag{1.5}$$

where $R \in M(S)$. Equation (1.5) is one of two equivalent parametrizations given on page 108 of [3]. It is customary to require that $\det(Y - R\tilde{N}) \neq 0$ in (1.5). It should be noted, however, that such "infinite" compensators are only some of the many unrealizable compensators that can be produced by inserting various values of R into the parametrization (1.5). A parametrization of all "possible sensitivities" that result from a stabilizing compensator can be produced by substituting (1.5) and (1.3) into (1.1) [3]. The result of this is that every "possible sensitivity" can be expressed as

$$S = \tilde{Y}\tilde{D} - NR\tilde{D}. \tag{1.6}$$

Every "possible weighted sensitivity" can now be expressed as

$$W_1 S W_2 = W_1 \tilde{Y}\tilde{D} W_2 - W_1 NR\tilde{D} W_2$$

where $R \in M(S)$. The sensitivity minimization problem has now been reduced to

$$\min_{R \in M(S)} \| W_1 \tilde{Y}\tilde{D} W_2 - W_1 NR\tilde{D} W_2 \|_\infty. \tag{1.7}$$

It is generally agreed that the minimization problem (1.7) should be considered solved if the suboptimal problem (1.8) is solved. Clearly, this could be done by repeating the solution of (1.8) for a sequence of values of ρ:

$$\text{Characterize } \{R \in M(S): \| W_1 \tilde{Y}\tilde{D} W_2 - W_1 NR\tilde{D} W_2 \|_\infty \leqq \rho\}. \tag{1.8}$$

It should also be noted that the solution to (1.8) is likely to be more useful than the solution to (1.7), since it may be used to allow "breathing room" on the sensitivity so that other constraints may be added.

Using (1.8) as a starting point, it can be shown that all "possible sensitivities" (1.6) that satisfy

$$\| W_1 S W_2 \|_\infty \leqq \rho$$

can be parametrized according to (1.9) using methods described in [20]:

$$S = [\theta_{11} G + \theta_{12}][\theta_{21} G + \theta_{22}]^{-1} \tag{1.9}$$

where $\|G\|_\infty \leqq 1$. This characterizes the set (1.8), in principle, because (1.6) can be backsolved for R. The corresponding compensator can then be computed from (1.5), or (1.1) can be backsolved for $C(s)$.

The idea of a two-frequency-scale rational matrix was introduced in [7]. Two-frequency-scale matrices are parameter-dependent rational matrices whose behavior becomes increasingly separated into high- and low-frequency categories as the parameter approaches a certain specified critical value. In this paper, the perturbation parameter is ε and the critical value is zero. A number of results concerning two- and multiple-frequency scale matrices are given in [7] and [8].

The objective of this paper is to show that if the data (Q, W_1, and W_2) of the sensitivity minimization problem all have two-frequency-scale behavior and satisfy certain regularity conditions, then approximate solutions can be generated by solving

two similar reduced-order subproblems. It is then shown, with considerably more difficulty, that the matrices θ_{ij} of (1.9) are two-frequency-scale if the data for the problem is two-frequency-scale.

The remainder of this paper is organized as follows. Section 2 gives background on two-frequency-scale systems and on the solution of (1.8). The background developed in § 2 is applied to give the main results of the paper in § 3.

2. Background.

2.1. Two-frequency-scale matrices. Consider a transfer matrix $H(s, \varepsilon)$ that is rational in s with coefficients that depend analytically on the parameter ε. In this paper, ε is assumed at all times to be a nonnegative real variable. The general behavior of such a transfer matrix is very complicated. The concept of a two-frequency-scale rational matrix narrows down the many possible types of behavior to a degree that allows many nontrivial results to be proved. The following three statements give an intuitive description of the two-frequency-scale property.

(1) All coefficients are real analytic in ε at $\varepsilon = 0$.

(2) The poles fall into two classes. Those in one class approach finite values as ε approaches zero. Those in the other class go to infinity as $K(\varepsilon)/\varepsilon$, where K approaches a nonzero constant as $\varepsilon \to 0$. K may have an algebraic singularity at $\varepsilon = 0$. Thus, the poles fall into two frequency scales—the low or "slow" frequency scale s and the high or "fast" frequency scale $p = \varepsilon s$. The slow poles approach finite limits in the s frequency scale as $\varepsilon \to 0$. The remaining fast poles approach finite, nonzero limits in the p frequency scale as $\varepsilon \to 0$. Thus, the polynomial $\varepsilon s^2 + 1$ is not allowable as a pole polynomial because the poles $\pm j/\sqrt{\varepsilon}$ neither go to finite values as $\varepsilon \to 0$, nor can they be expressed as $K(\varepsilon)/\varepsilon$, where $K(\varepsilon)$ has the required properties. An attempt gives $K(\varepsilon) = \pm j\sqrt{\varepsilon}$, which goes to zero as $\varepsilon \to 0$. An example of an allowable pole polynomial is $\varepsilon s + 1$. In this case, $K(\varepsilon) = -1$.

(3) Two-frequency-scale matrices must be rational and proper for all sufficiently small fixed values of ε. Furthermore, they must behave as rational, proper matrices in an asymptotic way in each frequency range. Thus, they behave as dynamical systems for all sufficiently small ε, and in the limit as $\varepsilon \to 0$ in each frequency range.

The preceding verbal descriptions of the two-frequency-scale matrices will now be made mathematically precise. We first let R_ε denote the ring of germs [9] of real-valued functions analytic in the real variable ε at $\varepsilon = 0$, and let F_ε be the field of quotients of R_ε. A rigorous definition of two-frequency-scale rational matrices can now be given.

DEFINITION 2.1. A matrix-valued function $H(s, \varepsilon)$ is two-frequency-scale if:

(1) $H(s, \varepsilon) \in F_\varepsilon(s)$; $H(s, \varepsilon)$ is proper in s.

(2) $H(s, 0) \in \mathbb{R}(s)$; $H(s, 0)$ is proper:

$$H\left(\frac{p}{\varepsilon}, \varepsilon\right)\Big|_{\varepsilon=0} \in \mathbb{R}(p), \qquad H\left(\frac{p}{\varepsilon}, \varepsilon\right)\Big|_{\varepsilon=0} \text{ is proper.}$$

(3) Each pole s_j of $H(s, \varepsilon)$ can be expanded in one of the following two ways:

$$\text{(2.1)} \quad \text{(A)} \qquad s_j(\varepsilon) = \sum_{i=0}^{\infty} a_{ji}\varepsilon^{i/q};$$

$$\text{(2.2)} \quad \text{(B)} \qquad s_j(\varepsilon) = \frac{1}{\varepsilon}\sum_{i=0}^{\infty} b_{ji}\varepsilon^{i/q}, \qquad b_{j0} \neq 0;$$

q is a positive integer that may depend on j.

It should be noted that part (B) of condition (3) above rules out denominators such as $\varepsilon^2 s+1$ and εs^2+1. Poles are forced to be "exactly" $O(1)$ or $O(1/\varepsilon)$ as $\varepsilon \to 0$. The positive integer q takes on nonunity values for denominators such as $s^2+\varepsilon=0$. The matrices computed in part (2) of Definition 2.1 are, for some purposes, approximations of the matrix $H(s, \varepsilon)$ for low and high frequency. These approximations are given a special notation for convenience in the following definition.

DEFINITION 2.2. Let $H(s, \varepsilon)$ be a two-frequency-scale rational matrix. Then the matrices H_S and H_F over $\mathbb{R}(s)$ and $\mathbb{R}(p)$, respectively, are given by

$$H_S(s) = H(s, 0), \tag{2.3}$$

$$H_F(p) = H\left(\frac{p}{\varepsilon}, \varepsilon\right)\Big|_{\varepsilon=0}. \tag{2.4}$$

Several senses in which H_S and H_F are approximations of H are made precise in [7] and [8].

The following theorem connects the frequency-domain concept of a two-frequency-scale rational matrix with the state-space idea of a singularly perturbed system, for which there is an extensive literature.

THEOREM 2.1. *A matrix $H(s, \varepsilon)$ is two-frequency-scale if and only if there exists a minimal realization of the form* (2.5) *for $H(s, \varepsilon)$ in which the matrices C_i, A_{ij}, B_j, and D are in $M(R_\varepsilon)$; and* $\det A_{22}(0) \neq 0$:

$$\dot{x}_1 = A_{11}x_1 + A_{12}x_2 + B_1 u, \tag{2.5a}$$

$$\dot{x}_2 = \frac{1}{\varepsilon} A_{21}x_1 + \frac{1}{\varepsilon} A_{22}x_2 + \frac{1}{\varepsilon} B_2 u, \tag{2.5b}$$

$$y = C_1 x_1 + C_2 x_2 + Du. \tag{2.5c}$$

Proof. For the proof see [10]. □

In words, Theorem 2.1 says that a matrix $H(s, \varepsilon)$ is two-frequency-scale if and only if it has a minimal analytic realization as a singularly perturbed system of differential equations. Showing that (2.5) has a two-frequency-scale transfer matrix is a straightforward computation involving system matrices [11]. Showing the converse requires a theorem for the realization of systems over rings [12]. It should be noted that "minimal" here means roughly "minimal for almost all sufficiently small ε."

It is clear that regularly perturbed systems, that is, systems of the form (2.5) with the dimension of x_2 equal to zero, form a subset of the singularly perturbed systems. Thus, any behavior that can occur in regularly perturbed systems is possible also for singularly perturbed systems. It is well known that a system depending on a single parameter that is completely controllable and observable for almost all values of the parameter may lose controllability and/or observability at discrete parameter values. In other words, the order of a minimal realization is a discontinuous function of the parameter in general. It turns out that a singularly perturbed system of form (2.5) can have poles that become asymptotically uncontrollable and/or unobservable as $\varepsilon \to 0$ in either or both of its frequency ranges. From the frequency-domain viewpoint, a two-frequency-scale rational matrix may have poles that do not appear in either $H_S(s)$ or $H_F(p)$. The concept of "lost poles," referring to poles that are "lost" in the limiting process as $\varepsilon \to 0$, is introduced to deal with this on a systematic basis. Lost poles are defined in [7] through the use of parameter-dependent system matrices. They are defined here, in an equivalent way, using state-space ideas. We first note that if the rational matrix $H(s, \varepsilon)$ has minimal realization (2.5), then the slow and fast descriptions

(2.3) and (2.4) can be expressed in terms of the blocks of (2.5) as follows:

$$H_S(s) = C_S(sI - A_S)^{-1}B_S + D_S, \tag{2.6}$$

$$H_F(p) = C_F(pI - A_F)^{-1}B_F + D_F \tag{2.7}$$

where

$$C_S = C_1 - C_2A_{22}^{-1}A_{21}|_{\varepsilon=0}, \tag{2.8a}$$

$$A_S = A_{11} - A_{12}A_{22}^{-1}A_{21}|_{\varepsilon=0}, \tag{2.8b}$$

$$B_S = B_1 - A_{12}A_{22}^{-1}B_2|_{\varepsilon=0}, \tag{2.8c}$$

$$D_S = D - C_2A_{22}^{-1}B_2|_{\varepsilon=0}, \tag{2.8d}$$

$$C_F = C_2(0), \tag{2.8e}$$

$$A_F = A_{22}(0), \tag{2.8f}$$

$$B_F = B_2(0), \tag{2.8g}$$

$$D_F = D(0). \tag{2.8h}$$

The subsystems (C_S, A_S, B_S, D_S) and (C_F, A_F, B_F, D_F) appear frequently in the time-domain singular perturbation literature (e.g., [13], [18]). Noting from the above discussion that the above-mentioned subsystems are not necessarily minimal, we can now define lost poles.

DEFINITION 2.3. Let $H(s, \varepsilon)$ be a two-frequency-scale rational matrix. Let (2.5) be a minimal-order realization whose existence is guaranteed by Theorem 2.1, and let C_S, A_S, B_S, and D_S be defined by (2.8). Then the lost slow poles of $H(s, \varepsilon)$ are those poles of the system (C_S, A_S, B_S, D_S) that are uncontrollable or unobservable. Similarly, the lost fast poles of $H(s, \varepsilon)$ are those poles of the system (C_F, A_F, B_F, D_F) that are uncontrollable or unobservable.

To apply coprime factorization theory to two-frequency-scale transfer matrices, an analogue of the set S is needed. The following definition of stable two-frequency-scale rational matrices is borrowed from [14].

DEFINITION 2.4. S_ε is the set of two-frequency-scale rational matrices whose poles satisfy the following condition. Slow poles (those that satisfy (2.1)) have Re $a_{j0} < 0$; fast poles (those that satisfy (2.2)) have Re $b_{j0} < 0$.

By Lemma 2.3 of [14], Definition 2.4 could also be stated as follows: a two-frequency-scale rational matrix $H(s, \varepsilon)$ is in S_ε if $H_S(s)$ and $H_F(p)$ are both stable and $H(s, \varepsilon)$ has no unstable lost poles. A portion of a theorem from [14] will now be extracted and paraphrased.

THEOREM 2.2. *Let $H(s, \varepsilon)$ be a two-frequency-scale rational matrix with no unstable lost poles. Then $H(s, \varepsilon)$ has right- and left-coprime factorizations over $M(S_\varepsilon)$ as follows*:

$$\begin{aligned} H(s, \varepsilon) &= N(s, \varepsilon)D(s, \varepsilon)^{-1} \\ &= \tilde{D}(s, \varepsilon)^{-1}\tilde{N}(s, \varepsilon). \end{aligned}$$

Furthermore, there exist X, Y, $\tilde{X}$, and $\tilde{Y}$ in $M(S_\varepsilon)$ such that

$$X(s, \varepsilon)N(s, \varepsilon) + Y(s, \varepsilon)D(s, \varepsilon) = I, \qquad \tilde{N}(s, \varepsilon)\tilde{X}(s, \varepsilon) + \tilde{D}(s, \varepsilon)\tilde{Y}(s, \varepsilon) = I.$$

Proof. For the proof see [14]. □

The proof of Theorem 2.2 is not obvious because the ring S_ε is not a Bezout domain. The proof in [14] is done through state-space methods. The results on

parametetrization of all stabilizing compensators described in § 1 now apply to two-frequency-scale rational matrices. It should be noted that "stable" means "in $M(S_\varepsilon)$" as pointed out in [14]. All allowable sensitivities can be generated by inserting any R in $M(S_\varepsilon)$ into (1.6). It should also be noted that the corresponding compensator need not be two-frequency-scale even if it is not infinite. This is not surprising in view of the remark following (1.5).

The theory of two-frequency-scale matrices described in this section is based on taking R_ε as the real germs at $\varepsilon = 0$. This leads to the matrices H_S and H_F having real coefficients. Although the results of [7], [8] were developed for complex germs, a nearly identical theory results if only real germs are considered.

The following result on block diagonalization of (2.5) is also needed [15].

THEOREM 2.3. *There exists a similarity transformation $T(\varepsilon) \in M(R_\varepsilon)$ that transforms system* (2.5) *into block-diagonal form as follows*:

$$T(\varepsilon)\begin{bmatrix} A_{11}(\varepsilon) & A_{21}(\varepsilon) \\ \frac{1}{\varepsilon}A_{21}(\varepsilon) & \frac{1}{\varepsilon}A_{22}(\varepsilon) \end{bmatrix} T(\varepsilon)^{-1} = A_D(\varepsilon)$$

$$= \begin{bmatrix} A_S + 0(\varepsilon) & 0 \\ 0 & \frac{1}{\varepsilon}A_F + 0(1) \end{bmatrix}$$

$$= \begin{bmatrix} A_1(\varepsilon) & 0 \\ 0 & \frac{1}{\varepsilon}A_2(\varepsilon) \end{bmatrix}.$$

Furthermore, $T(\varepsilon)$ and $T(\varepsilon)^{-1}$ have the following forms:

$$T(\varepsilon) = \begin{bmatrix} I + O(\varepsilon) & -\varepsilon A_{12}(0)A_F^{-1} + O(\varepsilon^2) \\ A_F^{-1}A_{21}(0) + O(\varepsilon) & I + O(\varepsilon) \end{bmatrix},$$

$$T(\varepsilon)^{-1} = \begin{bmatrix} I + O(\varepsilon) & \varepsilon A_{12}(0)A_F^{-1} + O(\varepsilon^2) \\ -A_F^{-1}A_{21}(0) + O(\varepsilon) & I + O(\varepsilon) \end{bmatrix}.$$

This theorem is the main tool of a number of papers on time-domain singular perturbation theory.

2.2. Solution of the function-disk problem. The state-space approach of [16] will be used to solve problem (1.8). A fairly algorithmic treatment will be discussed here since a theoretical discussion would be too lengthy to include. Let it suffice to mention that the approach, for the case $W_1 = W_2 = I$, involves looking at the graph space of the operator $S(s)$, treated as a mapping from $\tilde{D}^{-1}H_m^2$ to H_m^2, in a Krein space (indefinite inner product space) setting. The connection with (1.8) is that if ρ is normalized to 1, we require that the quantity within the norm symbol be a contraction mapping. A relationship between the graph spaces of contraction mappings and negative subspaces of the Krein space is then exploited.

A blanket assumption is that $Q(s)$ has no poles or zeros on the imaginary axis. This will guarantee that the process about to be described leads to a well-defined solution. If pure-imaginary poles or zeros of the plant are present, the algorithm will try to cancel them with pure-imaginary compensator zeros and poles.

We further assume that $Q(s)$ is square. Although the approach of [6] works in general, the state-space formulas of [16] are currently available only for the case when

$Q(s)$ is square and nonsingular at ∞. The first step is to form the matrices L, J', and J as follows:

$$L=\begin{bmatrix} W_1\tilde{Y} & W_1N \\ W_2^{-1}\tilde{D}^{-1} & 0 \end{bmatrix}, \quad J'=\begin{bmatrix} I & 0 \\ 0 & -\rho^2 I \end{bmatrix}, \quad J=\begin{bmatrix} I & 0 \\ 0 & -I \end{bmatrix} \tag{2.9}$$

where the matrices N, $\tilde{D}$, and $\tilde{Y}$ are as discussed in § 1. We now compute $\Theta(s)$ such that

$$L(s)=\Theta(s)F(s), \tag{2.10}$$ (1)

$$\Theta(-s)^*J'\Theta(s)=J, \tag{2.11}$$ (2)

F and F^{-1} are in $M(S)$. (2.12) (3)

Such a Θ, if it exists, is unique up to a J unitary constant right factor. If such a Θ can be found and the set (1.8) is nonvacuous, then all rational allowable sensitivities can be parametrized as

$$S(s)=[\theta_{11}(s)G(s)+\theta_{12}(s)][\theta_{21}(s)G(s)+\theta_{22}(s)]^{-1} \tag{2.13}$$

where $G(s)\in M(S)$ with $\|G\|_\infty \leqq 1$ and Θ is subdivided as

$$\Theta=\begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix}. \tag{2.14}$$

Thus, problem (1.8) has been converted to a J-inner-outer factorization problem.

The J-inner-outer factorization problem described above is solved in [16], with the unit disk as the domain of H^2 functions. This work is restated here for the right half-plane. The initial data for the problem is, by assumption, a minimal state-space realization for the matrix $L(s)$ defined in (2.9):

$$L(s)=C(sI-A)^{-1}B+D. \tag{2.15}$$

Viewing all matrices as operators, we now define

$$A^x=A-BD^{-1}C, \tag{2.16a}$$

$P=$ Reisz projection of A for the right half-plane restricted to its image (2.16b)

$$=\frac{1}{2\pi j}\int_\Gamma (sI-A)^{-1}\,ds,$$

where Γ encircles all right-half-plane eigenvalues of A.

$P^x=$ Riesz projection of A^x for the right half-plane restricted to its image, (2.16c)

$$C_-=J'^{-1}D^{-*}B^*\,|\,\mathrm{Im}\ P^{x*}, \tag{2.16d}$$

$$C_+=C\,|\,\mathrm{Im}\ P, \tag{2.16e}$$

$$B_-=P^*C^*J', \tag{2.16f}$$

$$B_+=P^xBD^{-1}, \tag{2.16g}$$

$$A_{p-}=-A^{x*}\,|\,\mathrm{Im}\ P^{x*}, \tag{2.16h}$$

$$A_{p+}=A\,|\,\mathrm{Im}\ P, \tag{2.16i}$$

$$S_1=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P^*(t+A^*)^{-1}C^*J'C(t-A)^{-1}P\,dt, \tag{2.16j}$$

$$S_2=\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P^x(t-A^x)^{-1}BD^{-1}J'^{-1}D^{-*}B^*(t+A^{x*})^{-1}P^{x*}\,dt, \tag{2.16k}$$

$$\mathscr{L} = \begin{bmatrix} P^* & S_1 \\ S_2 & P^x \end{bmatrix} \quad \text{where } \mathscr{L}\colon \operatorname{Im} P^{x*} + \operatorname{Im} P \to \operatorname{Im} P^* + \operatorname{Im} P^x. \tag{2.16l}$$

With these definitions, an expression for Θ (when it exists) can now be written down. It turns out that Θ exists if and only if $\mathscr{L}$ is invertible, and then

$$\Theta(s) = \left\{ I + [C_- C_+] \begin{bmatrix} s - A_{p-} & 0 \\ 0 & s - A_{p+} \end{bmatrix}^{-1} \mathscr{L}^{-1} \begin{bmatrix} B_- \\ B_+ \end{bmatrix} \right\} \Theta(\infty). \tag{2.17}$$

We note that Θ satisfying (2.10)-(2.12) exists if and only if $\mathscr{L}$ is invertible. If $\mathscr{L}$ is not invertible, it can be made invertible by making ρ slightly larger. It is possible to test whether the set (1.8) is nonvacuous by computing a generalized Pick matrix [6]. Another way to see if (1.8) is nonempty is to compute Θ using (2.17) and look at a sample output of the linear fractional map (2.13).

It is not obvious that (2.17) always produces the same result because there are arbitrary choices of the factorization (1.3)-(1.4) and of the state-space realization (2.15). It turns out, however, that the result is unique given Q, W_1, W_2, and $\Theta(\infty)$. This is because Θ is determined (up to the (J, J')-unitary constant $\Theta(\infty)$) by the behavior of Q, W_1, and W_2 at the right-half-plane poles and zeros of Q. This section ends with some facts from the theory of system matrices [11]. The system matrix notion provides a method of describing systems that is somewhat more general than state-space representation. This allows for a wider range of transformations that preserve the transfer matrix. The notation is defined by the relationship

$$\begin{bmatrix} T & U \\ -V & W \end{bmatrix} \sim VT^{-1}U + W. \tag{2.18}$$

It can be shown easily that system matrices can be left-multiplied by nonsingular lower-block-triangular factors from the left without changing the transfer matrix. A similar statement holds for right-multiplication by nonsingular upper-block-triangular factors. The following theorem, which concerns the special case when $W = I$, follows easily from formula (6.4) of [11] and several system matrix operations.

THEOREM 2.4. *Let H_i, for $i = 1, 2$, have system matrix descriptions*

$$\begin{bmatrix} T_i & U_i \\ -V_i & I \end{bmatrix} \sim H_i.$$

Then the product H_2H_1 has the following system matrix description:

$$\begin{bmatrix} -U_2V_1 & T_2 & U_2 \\ T_1 & 0 & U_1 \\ -V_1 & -V_2 & I \end{bmatrix} \sim H_2H_1. \tag{2.19}$$

3. A two-frequency-scale disk problem. We now consider the case where the plant Q of Fig. 1.1 and the weights W_i of (1.8) are two-frequency-scale and satisfy a set of regularity conditions. The regularity conditions essentially require that the solution to the problem defined by (1.8) is well-defined in each frequency scale, and that there are no singularities induced by unstable lost poles as discussed in the previous section. We start with two lemmas. The first concerns inversion of rational matrices in terms of a state-space form.

LEMMA 3.1 [19]. *Let $H(s)$ be a square rational matrix over an arbitrary field F, and suppose that $H(s)$ has a state-space realization*

$$H(s) = C(sI - A)^{-1}B + D$$

with D invertible. Then $H(s)^{-1}$ can be written

$$H(s)^{-1} = D^{-1} - D^{-1}C(sI - A^x)^{-1}BD^{-1}$$

where $A^x = A - BD^{-1}C$.

Proof. Lemma 3.1 can be proved easily using the Schur complement formula [17]. □

The next lemma concerns two-frequency-scale weighting matrices.

LEMMA 3.2. Let $W(s, \varepsilon)$ be a square two-frequency-scale matrix with no unstable lost poles. Suppose that $W_S(s)$ and $W_F(p)$ are both stable, and both have stable inverses. Then there exists $\varepsilon^* > 0$ such that for $\varepsilon \in [0, \varepsilon^*)$, $W(s, \varepsilon)$ is stable and has a stable inverse. Furthermore, $W(s, \varepsilon)^{-1}$ is two-frequency-scale and $[W(s, \varepsilon)^{-1}]_S = W_S(s)^{-1}$, and $[W(s, \varepsilon)^{-1}]_F = W_F(p)^{-1}$.

Proof. We first note that Theorem 3.5 of [7] shows that $W(s, \varepsilon)$ is stable for $\varepsilon \in [0, \varepsilon^*)$ for some $\varepsilon^* > 0$.

The corresponding fact for $W(s, \varepsilon)^{-1}$ will now be shown. Since $W_S(s)$ and $W_F(p)$ have stable inverses, they must be nonsingular at ∞. Lemma 2.5 of [14] shows that $W(s, \varepsilon)^{-1}$ is two-frequency-scale. The expressions for $[W(s, \varepsilon)^{-1}]_S$ and $[W(s, \varepsilon)^{-1}]_F$ follow by direct substitution. At this point, the only thing remaining to prove is that $W(s, \varepsilon)^{-1} \in S_\varepsilon$. Since $[W(s, \varepsilon)^{-1}]_S$ and $[W(s, \varepsilon)^{-1}]_F$ are both stable, we need to verify only that all lost poles of $W(s, \varepsilon)^{-1}$ are stable, by virtue of Lemma 2.3 of [14] as mentioned after Definition 2.4. It can be seen from Lemma 3.1 that the lost poles of $W(s, \varepsilon)^{-1}$ are the same as the lost poles of $W(s, \varepsilon)$. One way to show this is to use the Popov–Belevitch–Hautus rank tests [17, p. 136]. □

The following assumption is motivated by the lemma above.

Assumption A1. This assumption holds for $W_i(s, \varepsilon)$ with $i = 1, 2$. $W_i(s, \varepsilon)$ is two-frequency-scale and has no unstable lost poles in either frequency scale. Furthermore, the matrices $W_{iS}(s)$, $[W_{iS}(s)]^{-1}$, $W_{iF}(p)$, and $[W_{iF}(p)]^{-1}$ are all in $M(S)$. That is, they are stable and have stable inverses.

The assumption in A1 on lost poles is needed, for otherwise the stability of S would not be equivalent to the stability of $W_1 S W_2$. The need for the next assumption is clear, for if the plant had unstable lost poles, then the existence of a stabilizing, two-frequency-scale compensator would not be guaranteed, and Theorem 2.2 could not be applied.

Assumption A2. The plant $Q(s, \varepsilon)$ is square, is two-frequency-scale, and has no unstable lost poles in either frequency scale. Furthermore, $Q_S(s)$ and $Q_F(p)$ have no poles or zeros on the extended imaginary axis.

The objective of this section is to show that a full-order problem can be decomposed into two reduced-order subproblems, and that the solutions to the two subproblems can be combined to yield an approximate solution to the full-order problem. For the purpose of this paper, we will consider that solving problem (1.8) is equivalent to finding Θ satisfying (2.10)–(2.12) because of the parametrization (2.13)–(2.14).

FULL-ORDER PROBLEM. Given $Q(s, \varepsilon)$, $W_1(s, \varepsilon)$, and $W_2(s, \varepsilon)$ satisfying Assumptions A1 and A2, form left and right coprime factorizations of $Q(s, \varepsilon)$ whose existence is guaranteed by Theorem 2.2. Using the notation of this theorem, compute the blocks of the following matrix to produce $L(s, \varepsilon)$:

$$L(s, \varepsilon) = \begin{bmatrix} W_1(s, \varepsilon)\tilde{Y}(s, \varepsilon) & W_1(s, \varepsilon)N(s, \varepsilon) \\ W_2(s, \varepsilon)^{-1}\tilde{D}(s, \varepsilon)^{-1} & 0 \end{bmatrix}. \tag{3.1}$$

Find $\Theta(s, \varepsilon)$ satisfying:

(1) $L(s, \varepsilon) = \Theta(s, \varepsilon)F(s, \varepsilon)$;

(2) $\Theta(-s, \varepsilon)^* J' \Theta(s, \varepsilon) = J$;

(3) $F(s, \varepsilon)$ and $[F(s, \varepsilon)]^{-1}$ are in $M(S_\varepsilon)$;

(4) $$\Theta(\infty) = \begin{bmatrix} I & 0 \\ 0 & \rho^{-1} I \end{bmatrix}.$$

We now define the slow and fast subproblems.

SLOW SUBPROBLEM. Given Q, W_1, and W_2 satisfying Assumptions A1 and A2, compute $Q_S(s)$, $W_{1S}(s)$, and $W_{2S}(s)$, as well as left- and right-coprime factorizations of $Q_S(s)$. From the matrix

(3.2) $$L^S(s) = \begin{bmatrix} W_{1S}\tilde{Y}^S & W_{1S}N^S \\ W_{2S}^{-1}(\tilde{D}^S)^{-1} & 0 \end{bmatrix}.$$

Find $\Theta^S(s)$ such that:

(3.3a) $$L^S(s) = \Theta^S(s)F^S(s);$$

(3.3b) $$\Theta^S(-\bar{s})^* J' \Theta^S(s) = J;$$

(3.3c) $$F^S \text{ and } (F^S)^{-1} \text{ are in } M(S);$$

(3.3d) $$\Theta^S(\infty) = \operatorname{diag}(I, \rho^{-1} I).$$

FAST SUBPROBLEM. Given Q, W_1, and W_2 satisfying Assumptions A1 and A2, compute $Q_F(p)$, $W_{1F}(p)$, and $W_{2F}(p)$, as well as left- and right-coprime factorizations of $Q_F(p)$. Form the matrix

(3.4) $$L^F(p) = \begin{bmatrix} W_{1F}\tilde{Y}^F & W_{1F}N^F \\ W_{2F}^{-1}(\tilde{D}^F)^{-1} & 0 \end{bmatrix}.$$

Find $\Theta^F(p)$ satisfying obvious analogues of (3.3a)-(3.3d). That is, replace every occurrence of the superscript "S" by the superscript "F."

It should be noted that the matrices N^S, D^S, X^S, Y^S, $\tilde{N}^S$, $\tilde{D}^S$, $\tilde{X}^S$, and $\tilde{Y}^S$ are computed only from knowledge of Q_S, W_{1S}, and W_{2S}. Thus $L^S \neq L_S$ in general. A similar comment holds for the fast subproblem.

A theorem on compensator design can now be stated. In words, it says that an approximate solution of the full-order problem can be obtained by solving the two reduced-order subproblems. The approach is to show that a stabilizing compensator for the full-order system that "almost" satisfies the constraint of (1.8) can be produced by finding a pair of compensators from the slow and fast subproblem solutions and combining the results. We note that $\Theta_1(s)$ given by

(3.5) $$\Theta_1(s) = \Theta^S(s) \operatorname{diag}(I, \rho I)\Theta^F(0)$$

is the value of Θ for the slow subproblem when $\Theta(\infty)$ is set to the (J, J')-unitary value $\Theta^F(0)$.

THEOREM 3.1. *Let $Q(s, \varepsilon)$, $W_1(s, \varepsilon)$, and $W_2(s, \varepsilon)$ satisfy Assumptions* A1 *and* A2. *Suppose that solutions $\Theta^S(s)$ and $\Theta^F(p)$ to the two subproblems exist and that the set* (1.8) *is nonvacuous for both subproblems. Let $G^S(s)$ and $G^F(p)$ be matrices in $M(S)$ such that $G^S(\infty) = G^F(0)$ and the following inequalities are satisfied*:

(3.6a) $$\|G^S\|_\infty \leqq 1,$$

(3.6b) $$\|G^F\|_\infty \leqq 1.$$

Suppose that the linear fractional map (2.13) *is applied to* G^S *using* Θ_1 *defined in* (3.5) *to produce* S^S; *and that* (2.13) *is applied to* G^F *using* Θ^F *to produce* S^F. *Let* C^S *be the compensator generated by backsolving* (1.1) *for* C *when* S *is set equal to* S^S *and* Q *is set to* Q_S. *Let* C^F *be the same when* S *is set to* S^F *and* Q *is set to* Q_F. *Suppose the* C^S *and* C^F *produced this way are proper. Then the compensator* $\hat{C}(s, \varepsilon)$ *defined by*

$$\hat{C}(s, \varepsilon) = C^S(s) + C^F(\varepsilon s) - C^F(0) \tag{3.7}$$

stabilizes $Q(s, \varepsilon)$ *for all sufficiently small* ε. *Furthermore,*

$$\| W_1(s, \varepsilon) S(s, \varepsilon) W_2(s, \varepsilon) \|_\infty \leqq \rho + O(\varepsilon) \tag{3.8}$$

where

$$S(s, \varepsilon) = (I + Q(s, \varepsilon)\hat{C}(s, \varepsilon))^{-1}. \tag{3.9}$$

Proof. We first point out that C^S and C^F generated as above satisfy

$$C^S(\infty) = C^F(0). \tag{3.10}$$

This is true because $\Theta_1(\infty) = \Theta^F(0)$ and $G^S(\infty) = G^F(0)$. Thus, the linear fractional map (2.13) produces S^S and S^F that satisfy

$$S^S(\infty) = S^F(0).$$

Thus, since backsolving (1.1) gives a well-defined unique result, (3.10) holds. Clearly, $\hat{C}$ defined by (3.7) is two-frequency-scale and

$$\hat{C}_S(s) = C^S(s), \qquad \hat{C}_F(p) = C^F(p).$$

Since $\hat{C}_S$ stabilizes Q_S, $\hat{C}_F$ stabilizes Q_F, and there are no unstable lost poles of Q or $\hat{C}$ ($\hat{C}$ has no lost poles by construction), $\hat{C}(s, \varepsilon)$ stabilizes $Q(s, \varepsilon)$ by Corollary (3.1) of [7].

We now show that $S(s, \varepsilon)$ defined by (3.9) satisfies (3.8). $S(s, \varepsilon)$ is two-frequency-scale and has no pure imaginary lost poles (the closed-loop lost poles are also open-loop lost poles—see [7]). Theorem 4.2 of [8] shows that

$$\| S(s, \varepsilon) - \hat{S}(s, \varepsilon) \|_\infty = 0(\varepsilon)$$

where $\hat{S}(s, \varepsilon) = S_S(s) + S_F(\varepsilon s) - S_S(\infty)$.

By limiting arguments shown in the Appendix, we have

$$\| \hat{S}(s, \varepsilon) \|_\infty = \rho + O(\varepsilon).$$

We can now write

$$\begin{aligned} \| S(s, \varepsilon) \|_\infty &= \| S(s, \varepsilon) - \hat{S}(s, \varepsilon) + \hat{S}(s, \varepsilon) \|_\infty \\ &\leqq \| S(s, \varepsilon) - \hat{S}(s, \varepsilon) \|_\infty + \| \hat{S}(s, \varepsilon) \|_\infty \\ &= \rho + O(\varepsilon). \end{aligned}$$

□

The next theorem, which concerns the representor $\Theta(s, \varepsilon)$, shows that more can be said about this problem decomposition, and that results stronger than the "one-way" result of Theorem 3.1 may be possible. Also, it might ultimately prove more useful to approximate $\Theta(s, \varepsilon)$ by singular perturbations rather than by individual sensitivities satisfying (1.2). This is because it is difficult, in general, to express all of the physical constraints of a control problem as a single "disk" of the form (1.2) [5]. Thus, a possible approach is to first compute Θ and then incorporate other constraints by restricting the choice of the parameter G.

THEOREM 3.2. *Let $Q(s, \varepsilon)$, $W_1(s, \varepsilon)$, and $W_2(s, \varepsilon)$ satisfy Assumptions* A1 *and* A2. *Suppose that solutions $\Theta^S(s)$ and $\Theta^F(p)$, to the slow and fast subproblems, respectively, exist. Then a solution $\Theta(s, \varepsilon)$ to the full-order problem exists for all sufficiently small ε. Furthermore, $\Theta(s, \varepsilon)$ is two-frequency-scale and*

$$\Theta_S(s) = \Theta^S(s)\ \mathrm{diag}\,(I, \rho I)\Theta^F(0), \qquad \Theta_F(p) = \Theta^F(p).$$

Proof. We first note that $L(s, \varepsilon)$ given by (3.1) is two-frequency-scale. This follows from Assumptions A1 and A2, and from Lemma 3.2. Since $L(s, \varepsilon)$ is two-frequency-scale, it has a state-space realization of the form (2.5). The matrix $D(0)$ in (2.8h) is nonsingular because all of the following matrices are nonsingular at ∞: $W_{1F}(p)$, $N_F(p)$, $W_{2F}(p)^{-1}$, and $\tilde{D}_F(p)^{-1}$. This shows that Lemma 3.1 applies to $L(s, \varepsilon)$. The next step is to insert the state-space realization for $L(s, \varepsilon)$ into (2.16a)-(l) and generate the corresponding $\Theta(s, \varepsilon)$ using (2.17). Formula (2.17) is not quite in state-space form: the multiplication of $\mathscr{L}^{-1}$, $[B_-^T B_+^T]^T$, and $\Theta(\infty)$ would have to be carried out. The term multiplying $\Theta(\infty)$, however, fits naturally into the system matrix format. Formula (2.17) gives a minimal state-space realization of Θ if the appropriate multiplications are carried out. Thus, the domain restrictions were made so the state-space is as small as possible. It turns out that it is easier to use some unrestricted operators here. The validity of the following system matrix representation is easily seen from (2.17):

$$\Theta(s, \varepsilon)\Theta(\infty, \varepsilon)^{-1} \sim \begin{bmatrix} P^*(s + A^{x*})P^{x*} & P^*S_1(s - A)P & P^*C^*J' \\ P^xS_2(s + A^{x*})P^{x*} & P^x(s - A)P & P^xBD^{-1} \\ -J'^{-1}D^{-*}B^*P^{x*} & -CP & I \end{bmatrix} \tag{3.11}$$

where (C, A, B, D) is a state-space realization of $L(s, \varepsilon)$ of the form (2.5). Thus,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ \dfrac{1}{\varepsilon} A_{21} & \dfrac{1}{\varepsilon} A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ \dfrac{1}{\varepsilon} B_2 \end{bmatrix}, \quad C = [C_1 \quad C_2]$$

with $A_{22}(0)$ nonsingular. The remainder of the blocks of (3.11) are defined in (2.16). System matrix operations are now performed on (3.11) to show that $\Theta(s, \varepsilon)$ is two-frequency-scale and to compute $\Theta_S(s)$ and $\Theta_F(p)$.

As a first step toward this end, we note that the projection matrices $P(\varepsilon)$ and $P(\varepsilon)^x$ are analytic at $\varepsilon = 0$; furthermore, they have a special structure in the limit as ε approaches zero. Using the notation of Theorem 2.3, and letting $\Gamma(\varepsilon)$ be a curve that encircles all of the right-half-plane eigenvalues of A, we have

$$\begin{aligned} P(\varepsilon) &= \int_{\Gamma(\varepsilon)} (s - A)^{-1}\, ds \\ &= T(\varepsilon)^{-1} \int_{\Gamma(\varepsilon)} T(\varepsilon)(s - A)^{-1} T(\varepsilon)^{-1}\, ds\, T(\varepsilon) \\ &= T(\varepsilon)^{-1} \int_{\Gamma(\varepsilon)} (s - A_D)^{-1}\, ds\, T(\varepsilon) \\ &= T(\varepsilon)^{-1} P_D(\varepsilon) T(\varepsilon) \end{aligned} \tag{3.12}$$

where

$$\begin{aligned} &P_D(\varepsilon) = \mathrm{diag}\,(P_S(\varepsilon), P_F(\varepsilon)), \\ &P_S(\varepsilon) = \text{RHP Riesz projection for } A_1(\varepsilon), \\ &P_F(\varepsilon) = \text{RHP Riesz projection for } A_2(\varepsilon). \end{aligned}$$

P_S and P_F are analytic in ε by virtue of (2.16b). The block expressions for $T(\varepsilon)$ and $T(\varepsilon)^{-1}$ can be inserted into (3.12) to show that

$$P(\varepsilon)=\begin{bmatrix} P_S+O(\varepsilon) & \varepsilon[-P_SA_{12}A_{22}^{-1}+A_{12}A_{22}^{-1}P_F]+O(\varepsilon^2) \\ -A_{22}^{-1}A_{21}P_S+P_FA_{22}^{-1}A_{21}+O(\varepsilon) & P_F+O(\varepsilon) \end{bmatrix}. \tag{3.13}$$

All quantities A_{ij} on the right-hand side of (3.13) are evaluated at $\varepsilon=0$. Formulas similar to (3.12) and (3.13) for P^x can be written as follows. The analogue of (3.12) is

$$P^x(\varepsilon)=T^x(\varepsilon)^{-1}P_D^x(\varepsilon)T^x(\varepsilon) \tag{3.14}$$

where $T^x(\varepsilon)$ block diagonalizes A^x as in Theorem 2.3 and $P_D^x(\varepsilon)$ is the (block diagonal) right-half-plane Riesz projection for A_D^x.

We can now take advantage of the singular perturbation structure of (3.11) by replacing $P(\varepsilon)$ with (3.12) and replacing $P^x(\varepsilon)$ with (3.14). After doing this, the system matrix is simplified by the following steps:

(1) Left-multiplying the first block row by T^{-*};
(2) Left-multiplying the second block row by T^*;
(3) Right-multiplying the first block column by $(T^x)^{-*}$;
(4) Right-multiplying the second block column by T^{-1}.

These operations transform (3.11) to

$$\begin{bmatrix} P_D^*T^{-*}(s+A^{x*})P_D^{x*} & P_D^*T^{-*}S_1(s-A)T^{-1}P_D & P_D^*T^{-*}C^*J' \\ P_D^xT^xS_2(s+A^{x*})T^{x*}P_D^{x*} & P_D^xT^x(s-A)T^{-1}P_D & P_D^xT^xBD^{-1} \\ J'^{-1}D^{-*}B^*T^{x*}P_D^{x*} & CT^{-1}P_D & I \end{bmatrix}. \tag{3.15}$$

Matrix (3.15) can be further simplified by inserting the identity terms $T^{x*}(T^x)^{-*}$, $T^{-1}T$, $T^{x*}(T^x)^{-*}$, and $T^{-1}T$ into the (1, 1), (1, 2), (2, 1), and (2, 2) blocks of (3.15), respectively. This substitution reduces (3.15) to the following form:

$$\begin{bmatrix} P_D^*(T^{-*}T^{x*})(s+A_D^{x*})P_D^{x*} & P_D^*S_{1D}(s-A_D)P_D & P_D^*C_D^*J' \\ P_D^xS_{2D}(s+A_D^{x*})P_D^{x*} & P_D^x(T^xT^{-1})(s-A_D)P_D & P_D^xB_D \\ -J'^{-1}B_D^*P_D^{x*} & -C_DP_D & I \end{bmatrix}. \tag{3.16}$$

The following new notation has been introduced:

$$S_{1D}=T^{-*}S_1T^{-1}, \tag{3.17a}$$

$$S_{2D}=T^xS_2(T^x)^{-1}, \tag{3.17b}$$

$$B_D=T^xBD^{-1}, \tag{3.17c}$$

$$C_D=CT^{-1}. \tag{3.17d}$$

The objective here is to express all blocks appearing in (3.16) in terms of quantities associated with the slow and fast subsystems. We define

$$S_{1S}=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}P_S^*(t+A_S^*)^{-1}C_S^*J'C_S(t-A_S)^{-1}P_S\,dt, \tag{3.18a}$$

$$S_{1F}=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}P_F^*(t+A_F^*)^{-1}C_F^*J'C_F(t-A_F)^{-1}P_F\,dt, \tag{3.18b}$$

$$S_{2S}=\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}P_S^x(t-A_S^x)^{-1}B_SD_S^{-1}J'^{-1}D_S^{-*}B_S^*(t+A_S^{x*})^{-1}P_S^{x*}\,dt, \tag{3.18c}$$

$$S_{2F}=\frac{1}{2\pi j}\int_{-j\infty}^{j\infty}P_F^x(t-A_F^x)^{-1}B_FD_F^{-1}J'^{-1}D_F^{-*}B_F(t+A_F^{x*})^{-1}P_F^{x*}\,dt. \tag{3.18d}$$

It should be noted here that the quantities S_{1S} and S_{2S} would be calculated in solving the slow subproblem if the *particular* factorization

$$N^S(s) = N_S(s), \quad \tilde{D}^S(s) = \tilde{D}_S(s), \quad \tilde{Y}^S(s) = \tilde{Y}_S(s)$$

had been used in forming $L^S(s)$, in (3.2), and if the *particular* realization (C_S, A_S, B_S, D_S) had been used to compute $\Theta^S(s)$. Thus, the expressions (3.18a, c) *could* arise in solving the slow subproblem. A similar statement concerning S_{1F} and S_{2F} holds. We wish to express (3.17a–d), and thus the entire system matrix (3.16), in terms of quantities of this type. It can be shown through rather extensive calculation that (3.17a–d) take the following forms:

$$S_{1D} = \begin{bmatrix} S_{1S} + O(\varepsilon) & \varepsilon G_1 + O(\varepsilon^2) \\ \varepsilon G_1^* + O(\varepsilon^2) & \varepsilon S_{1F} + O(\varepsilon^2) \end{bmatrix} \tag{3.19a}$$

where

$$G_1 = P_S^* C_S J' C_F^* A_F^{-*} P_F, \tag{3.19b}$$

$$S_{2D} = \begin{bmatrix} S_{2S} + O(\varepsilon) & H_1 + O(\varepsilon) \\ H_1^* + O(\varepsilon) & \dfrac{1}{\varepsilon} S_{2F} + O(1) \end{bmatrix} \tag{3.19c}$$

where

$$H_1 = -P_S^x B_S D_S^{-1} J'^{-1} D_F^{-*} B_F (A_F^x)^{-*} P_F^{x*}, \tag{3.19d}$$

$$B_D = \begin{bmatrix} B_S D_S^{-1} + O(\varepsilon) \\ \dfrac{1}{\varepsilon} B_F D_F^{-1} + O(1) \end{bmatrix}, \tag{3.19e}$$

$$C_D = [C_S + O(\varepsilon) \quad C_F + O(\varepsilon)]. \tag{3.19f}$$

Results on gramians (solutions of Lyapunov equations) already exist in the time-domain literature (e.g., [18]). The elegance of the complex variable proof, however, warrants the inclusion of a sample computation in this paper. A derivation of (3.19a) appears in the Appendix along with the computations for (3.19e).

We are now in a position to compute $\Theta_F(p)$. To do this, we must substitute $p = \varepsilon s$ in (3.16), make the resulting system matrix analytic in ε by multiplying appropriate block columns by ε, and set ε to zero. If the upper-left block of the resulting (ε-independent) system matrix is nonsingular, then it is a system matrix representation of $\Theta_F(p)$. We now have five block rows and five block columns. It turns out that if we make the substitution and multiply the first three block columns and the fourth block row by ε, the the resulting matrix is analytic. When this is done, the following system matrix results:

$$\begin{bmatrix} P_D^* \begin{bmatrix} P & -K_L(p + A_F^{x*}) \\ 0 & p + A_F^{x*} \end{bmatrix} P_D^{x*} & P_D^* \begin{bmatrix} S_{1S} p & G_1(p - A_F) \\ 0 & S_{1F}(p - A_F) \end{bmatrix} P_D & P_D^* \begin{bmatrix} C_S^* \\ C_F^* \end{bmatrix} J' \\ P_D^x \begin{bmatrix} S_{2S} p & H_1(p + A_F^{x*}) \\ 0 & S_{2F}(p + A_F^{x*}) \end{bmatrix} P_D^{x*} & P_D^x \begin{bmatrix} P & K_U(p - A_F) \\ 0 & p - A_F \end{bmatrix} P_D & P_D^x \begin{bmatrix} B_S D_S^{-1} \\ B_F D_F^{-1} \end{bmatrix} \\ -J'^{-1}[0 \quad D_F^{-*} B_F^*] P_D^{x*} & -[0 \quad C_F] P_D & I \end{bmatrix} \tag{3.20}$$

G_1 and H_1 are defined by (3.19b) and (3.19d), respectively. The product of transformations $T^x T^{-1}$ clearly has the form (3.21)

$$T^x T^{-1} = \begin{bmatrix} I + O(\varepsilon) & \varepsilon K_U + O(\varepsilon^2) \\ K_L + O(\varepsilon) & I + O(\varepsilon) \end{bmatrix} \tag{3.21a}$$

where

$$K_L = A_F^{-x} A_{21}^x(0) - A_F^{-1} A_{21}(0), \tag{3.21b}$$

$$K_U = A_{12}(0) A_F^{-1} - A_{12}^x(0)(A_F^x)^{-1}. \tag{3.21c}$$

When the remaining multiplications are written out in (3.20) it can be seen that rearrangement of the first four block rows and columns yields the following structure:

$$\begin{bmatrix} * & * & * & * & * \\ * & * & * & * & * \\ ? & ? & 0 & 0 & ? \\ ? & ? & 0 & 0 & ? \\ ? & ? & 0 & 0 & I \end{bmatrix}.$$

The *'s and ?'s represent possibly nonzero entries. A system matrix of this form may be reduced, by deleting rows and columns, leaving only the ?'s.

After this is sorted out, we get

$$\Theta(p/\varepsilon)\cdot\Theta(\infty)^{-1}\big|_{\varepsilon=0} \sim \begin{bmatrix} P_F^*(p + A_F^{x*})P_F^{x*} & P_F^* S_{1F}(p - A_F)P_F & P_F^* C_F^* J' \\ P_F^x S_{2F}(p + A_F^{x*})P_F^{x*} & P_F^x(p - A_F)P_F & P_F^x B_F D_F^{-1} \\ J'^{-1} D_F^{-*} B_F^* P_F^{x*} & C_F P_F & I \end{bmatrix}. \tag{3.22}$$

Clearly, the blocks of (3.22) could be the result of a computation of $\Theta^F(p)$ (i.e., the fast subproblem computation) if a particular factorization for the plant and a particular state-space realization for $L^F(p)$ were chosen. Thus,

$$\Theta_F(p) = \Theta\left(\frac{p}{\varepsilon}, \varepsilon\right)\Big|_{\varepsilon=0} = \Theta^F(p). \tag{3.23}$$

The computation of $\Theta_S(s)$ follows along a similar line. If system matrix (3.16) is viewed as having five block rows and five block columns (after the substitutions (3.19) are made), it can be made analytic in ε by multiplying the third block row and the second block column by ε. After setting ε to zero, (3.24) is obtained:

(3.24)

$$\begin{bmatrix} P_D^* \begin{bmatrix} s + A_S^{x*} & K_L^* A_F^{x*} \\ 0 & A_F^{x*} \end{bmatrix} P_D^{x*} & P_D^* \begin{bmatrix} S_{1S}(s - A_S) & -G_1 A_F \\ 0 & -S_{1F} A_F \end{bmatrix} P_D & P_D^* \begin{bmatrix} C_S^* \\ C_F^* \end{bmatrix} J' \\ P_D^x \begin{bmatrix} S_{2S}(s + A_F^{x*}) & H_1 A_F^{x*} \\ 0 & S_{2F} A_F^{x*} \end{bmatrix} P_D^{x*} & P_D^x \begin{bmatrix} s - A_S & -K_U A_F \\ 0 & -A_F \end{bmatrix} P_D & P_D^x \begin{bmatrix} B_S D_S^{-1} \\ B_F D^{-1} \end{bmatrix} \\ -J'^{-1}[D_S^{-*} B_S^* \quad D^{-*} B_F^*] P_D^{x*} & -[C_S \quad C_F] P_D & I \end{bmatrix}.$$

The rows and columns of (3.24) can be interchanged to place all of the zeros in the same block. This leaves (3.25), which is similar in form to (2.25):

$$\begin{bmatrix} X & T_2 & U_2 \\ T_1 & 0 & U_1 \\ -V_1 & -V_2 & I \end{bmatrix} \tag{3.25}$$

where

$$X = \begin{bmatrix} P_S^* K_L^* A_F^{x*} P_F^{x*} - P_S^* G_1 A_F P_F \\ P_S^x H_1 A_F^{x*} P_F^{x*} - P_S^x K_U A_F P_F \end{bmatrix},$$

$$T_2=\begin{bmatrix} P_S^*(s+A_S^{x*})P_S^{x*} & P_S^*S_{1S}(s-A_S)P_S \\ P_S^xS_{2S}(s+A_S^{x*})P_S^{x*} & P_S^x(s-A_S)P_S \end{bmatrix}, \qquad U_2=\begin{bmatrix} P_S^*C_S^* & J' \\ P_S^*B_S & D_S^{-1} \end{bmatrix},$$

$$T_1=\begin{bmatrix} P_F^*A_F^{x*}P_F^{x*}-P_F^*S_{1F}A_FP_F \\ P_F^xS_{2F}A_F^{x*}P_F^{x*}-P_F^xA_FP_F \end{bmatrix}, \qquad U_1=\begin{bmatrix} P_F^*C_F^*J' \\ P_F^*B_FD^{-1} \end{bmatrix},$$

$$V_1=[-J'^{-1}D^{-*}B_F^*P_F^{x*} \quad C_FP_F], \qquad V_2=[-J'^{-1}D_S^{-*}B_S^*P_S^{x*} \quad C_SP_S].$$

A calculation shows that indeed, X is equal to $-U_2V_1$, making the match to (2.25) complete. It is easily seen that

$$\begin{bmatrix} T_1 & U_1 \\ -V_1 & I \end{bmatrix} \sim \Theta^F(0)\ \mathrm{diag}\,(I, \rho I), \tag{3.26}$$

$$\begin{bmatrix} T_2 & U_2 \\ -V_2 & I \end{bmatrix} \sim \Theta^S(s)\ \mathrm{diag}\,(I, \rho). \tag{3.27}$$

Thus, from Theorem 2.4, we see that

$$\Theta(s, \varepsilon)|_{\varepsilon=0}\ \mathrm{diag}\,(I, \rho I)=\Theta^S(s)\cdot \mathrm{diag}\,(I, \rho I)\cdot \Theta^F(0)\cdot \mathrm{diag}\,(I, \rho I).$$

We have shown that $\Theta(s, \varepsilon)$ obeys parts (1) and (2) of Definition 2.1. Part (3) follows immediately from (2.17). The poles of $\Theta(s, \varepsilon)$ fall into two categories: they are eigenvalues of either $-A^{x*}$ or A. In either case, the poles obey (2.1) or (2.2) because both $-A^{x*}$ and A have the "singularly perturbed" form. □

Appendix.

Proof of (3.19*a*).

$$\begin{aligned} T^{-*}S_1T^{-1} &= -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} T^{-*}P^*(t+A^*)^{-1}C^*J'C(t-A)^{-1}PT^{-1}\,dt \\ &= -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} T^{-*}P^*T^*T^{-*}(t+A^*)^{-1}T^*T^{-*}C^*J'CT^{-1} \\ &\quad \times T(t-A)^{-1}T^{-1}TPT^{-1}\,dt \\ &= -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_D^*(t+A_D^*)^{-1}T^{-*}C^*J'CT^{-1}(t-A_D)^{-1}P_D\,dt \\ &= \begin{bmatrix} X_{11} & X_{12} \\ X_{12}^* & X_{22} \end{bmatrix}+\begin{bmatrix} E_{11} & E_{12} \\ E_{12}^* & E_{22} \end{bmatrix} \end{aligned} \tag{A1}$$

where

$$X_{11}=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_S^*(t+A_1^*)^{-1}C_S^*J'C_S(t-A_1)^{-1}P_S\,dt, \tag{A2}$$

$$X_{12}=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_S^*(t+A_1^*)^{-1}C_S^*J'C_F\left(t-\frac{1}{\varepsilon}A_2\right)^{-1}P_F\,dt, \tag{A3}$$

$$X_{22}=-\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_F^*\left(t+\frac{1}{\varepsilon}A_2^*\right)^{-1}C_F^*J'C_F\left(t-\frac{1}{\varepsilon}A_2\right)^{-1}P_F\,dt, \tag{A4}$$

and $A_D=\mathrm{diag}\,(A_1, (1/\varepsilon)A_2)$. Since

$$T^{-*}C^*J'CT^{-1}=\begin{bmatrix} C_S^*J'C_S & C_S^*J'C_F \\ C_F^*J'C_S & C_F^*J'C_F \end{bmatrix}+O(\varepsilon),$$

each E_{ij} will be one order of ε smaller than the corresponding X_{ij}. In other words, the X_{ij}'s form a first-order approximation of the gramian (A1). We first note that $X_{11}(0) = S_{1S}$ because $A_1 = A_S + O(\varepsilon)$. X_{22} can be computed by changing variables. Let $p = \varepsilon t$

$$X_{22} = -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_F^*(p + A_2^*)^{-1} C_F^* J' C_F (p - A_2)^{-1} P_F \cdot \varepsilon^2 \frac{dp}{\varepsilon}$$

$$= \varepsilon \cdot S_{1F} + O(\varepsilon^2),$$

since $A_2 = A_F + O(\varepsilon)$.

The cross-term X_{12} can be computed by contour integration:

$$X_{12}^* = -\frac{1}{2\pi j}\int_{-j\infty}^{j\infty} P_F^*\left(t + \frac{1}{\varepsilon} A_2^*\right)^{-1} C_F^* J' C_S (t - A_1)^{-1} P_S \, dt.$$

Since the integrand falls off as $O(1/t^2)$ for large t, this integral can be replaced by a large D-shaped contour integral. This is diagrammed in Fig. A1. Thus, if the radius of Γ is assumed of order larger than $1/\varepsilon$,

$$X_{12}^* = -\frac{1}{2\pi j}\int_{\Gamma} P_F^*\left(t + \frac{1}{\varepsilon} A_2^*\right)^{-1} C_F^* J' C_S (t - A_1)^{-1} P_S \, dt.$$

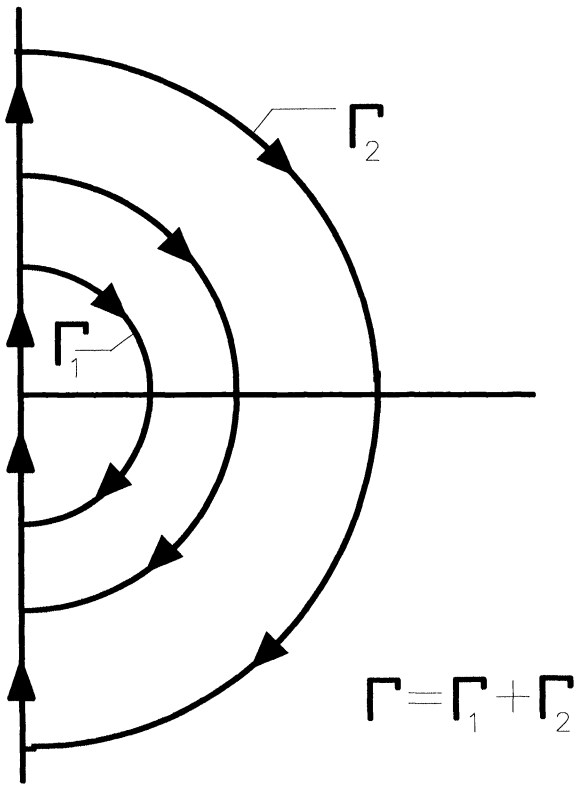

FIG. A1

We now note that

$$P_F^*\left(t+\frac{1}{\varepsilon}A_2^*\right)^{-1} = P_F^*\left(t+\frac{1}{\varepsilon}A_2^*\right)^{-1}P_F^*$$

is stable for all sufficiently small ε, and all the poles of $(t-A_1)^{-1}$ remain finite as $\varepsilon \to 0$. Choose ε^* so that Γ_1 encircles all the poles of $(t-A_1)^{-1}$ for all $\varepsilon < \varepsilon^*$. We can express the contour Γ as $\Gamma = \Gamma_1 + \Gamma_2$. The integral, however, is zero over Γ_2 because the integrand is analytic inside Γ_2 (at least for small ε). Thus,

$$\begin{aligned} X_{12}^* &= -\frac{1}{2\pi j}\int_{\Gamma_1} P_F^*\left(t+\frac{1}{\varepsilon}A_2^*\right)^{-1} C_F^* J' C_S (t-A_1)^{-1} P_S \, dt \\ &= -\frac{\varepsilon}{2\pi j}\int_{\Gamma_1} P_F^*(\varepsilon t + A_2^*)^{-1} C_F^* J' C_S (t-A_1)^{-1} P_S \, dt \\ &= -\frac{\varepsilon}{2\pi j}\int_{\Gamma_1} P_F^* A_2^{-*} C_F^* J' C_S (t-A_1)^{-1} P_S \, dt + \text{higher-order terms in } \varepsilon \\ &= \varepsilon P_F^* A_F^{-*} C_F^* J' C_S P_S \end{aligned}$$

because $-(1/2\pi j)\int_{\Gamma_1}(t-A_1)^{-1}\,dt = P_S$. Therefore, $X_{12} = \varepsilon P_S^* C_S^* J' C_F A_F^{-1} P_F + O(\varepsilon^2)$ and $G_1 = P_S^* C_S^* J' C_F A_F^{-1} P_F$. The proof of (3.19c) goes in a similar manner. □

Proof of (3.19e). We first note that a system of the form (2.5) has corresponding A^x given by

$$A^x = \begin{bmatrix} A_{11} - B_1 D^{-1} C_1 & A_{12} - B_1 D^{-1} C_2 \\ \dfrac{A_{21} - B_2 D^{-1} C_1}{\varepsilon} & \dfrac{A_{22} - B_2 D^{-1} C_2}{\varepsilon} \end{bmatrix}.$$

The transformation T^x of A^x to block diagonal form given by Theorem 2.3 can now be expressed as follows:

$$T^x = \begin{bmatrix} I & -\varepsilon(A_{12} - B_1 D^{-1} C_2) & (A_{22} - B_2 D^{-1} C_2)^{-1} \\ (A_{22} - B_2 D^{-1} C_2)^{-1} & (A_{21} - B_2 D^{-1} C_1) & I \end{bmatrix} + \begin{bmatrix} O(\varepsilon) & O(\varepsilon^2) \\ O(\varepsilon) & O(\varepsilon) \end{bmatrix}.$$

B has the form

$$B = \begin{bmatrix} B_1 \\ (1/\varepsilon) B_2 \end{bmatrix}.$$

Let

$$T^x B D^{-1} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}.$$

Clearly,

$$\begin{aligned} X_2 &= \frac{B_2 D^{-1}}{\varepsilon} + O(1) \\ &= \frac{B_F D_F^{-1}}{\varepsilon} + O(1). \end{aligned}$$

The computation of the limiting behavior of X_1 is more complicated. Multiplying T^x, B, and D^{-1}, we see that

$$X_1 = [B_1 - (A_{12} - B_1 D^{-1} C_2)(A_{22} - B_2 D^{-1} C_2)^{-1} B_2] D^{-1} + O(\varepsilon).$$

We can now expand $B_S D_S^{-1}$ (the right-hand side of the following expression is evaluated at $\varepsilon = 0$):

$$B_S D_S^{-1} = (B_1 - A_{12}A_{22}^{-1}B_2)(D - C_2 A_{22}^{-1}B_2)^{-1}.$$

The Schur complement formula gives us

$$\begin{aligned} D_S^{-1} &= (D - C_2 A_{22}^{-1}B_2)^{-1} \\ &= D^{-1} + D^{-1}C_2(-B_2 D^{-1}C_2 + A_{22})^{-1}B_2 D^{-1}. \end{aligned}$$

Thus, $B_S D_S^{-1} = (B_1 - A_{12}A_{22}^{-1}B_2)[I + D^{-1}C_2(A_{22} - B_2 D^{-1}C_2)^{-1}B_2]D^{-1}$. After multiplying out and collecting the term $(A_{22} - B_2 D^{-1}C_2)^{-1}B_2$, we get

$$\begin{aligned} B_S D_S^{-1} &= \{B_1 + [-A_{12} + A_{12}A_{22}^{-1}B_2 D^{-1}C_2 + B_1 D^{-1}C_2 - A_{12}A_{22}^{-1}B_2 D^{-1}C_2] \\ &\quad \times (A_{22} - B_2 D^{-1}C_2)^{-1}B_2\} D^{-1} \\ &= [B_1 - (A_{12} - B_1 D^{-1}C_2)(A_{22} - B_2 D^{-1}C_2)^{-1}B_2]D^{-1} \\ &= X_1 + O(\varepsilon). \end{aligned}$$

This proves (3.19e). □

Completion of the proof of Theorem 3.1. We start with the following lemma.

LEMMA A1. *Let $A(z)$ and $B(z)$ be matrices real analytic at $z = 0$, and let r be real with $0 < r \leqq \frac{1}{2}$. Also, let $A(0) \neq 0$ and $B(0) = 0$, and let*

$$\|A(j\xi)\|_2 \leqq \rho \quad \textit{for all } \xi \in \mathbb{R}.$$

Then $\|A(j\varepsilon^r) + B(j\varepsilon^{1-r})\|_2 \leqq \rho + O(\varepsilon)$ $(\varepsilon \to 0)$. Furthermore, the implicit bounding constant is independent of r.

Proof. We assume, without loss of generality, that $\|A(0)\|_2 = \rho$, since if $\|A(0)\|_2 \leqq \rho$, then the lemma follows immediately. We note that $\|\cdot\|_2$ is a continuous function of its argument. Thus,

$$\|A(j\varepsilon^r) + B(j\varepsilon^{1-r})\|_2 = \|A(0)\|_2 + f(\varepsilon) = \rho + f(\varepsilon)$$

where $f(\varepsilon)$ is continuous at zero and $f(0) = 0$. By elementary properties of the matrix norm, there exists a vector function $x(\varepsilon)$ such that $\|x(\varepsilon)\|_2 = 1$ and

$$\|[A(j\varepsilon^r) + B(j\varepsilon^{1-r})]x(\varepsilon)\|_2 = \rho + f(\varepsilon). \tag{A5}$$

Note that $\|A(j\varepsilon^r)x(\varepsilon)\|_2$ is bounded away from zero for small ε, since (A5) implies

$$\begin{aligned} \|A(j\varepsilon^r)x(\varepsilon)\|_2 &\geqq \rho + f(\varepsilon) - \|B(j\varepsilon^{1-r})x(\varepsilon)\|_2 \\ &\to \rho \quad \text{as } \varepsilon \to 0. \end{aligned}$$

The left-hand side of (A5) can be rewritten as

$$\begin{aligned} \|[A(j\varepsilon^r) + B(j\varepsilon^{1-r})]x(\varepsilon)\|_2 &= \{x(\varepsilon)^*[A(j\varepsilon^r) + B(j\varepsilon^{1-r})]^*[A(j\varepsilon^r) \\ &\quad + B(j\varepsilon^{1-r})]x(\varepsilon)\}^{1/2} \\ &= \{x(\varepsilon)^*[A^T(-j\varepsilon^r)A(j\varepsilon^r) + B^T(-j\varepsilon^{1-r})A(j\varepsilon^r) \\ &\quad + A^T(-j\varepsilon^r)B(j\varepsilon^{1-r}) + B^T(-j\varepsilon^{1-r})B(j\varepsilon^{1-r})]x(\varepsilon)\}^{1/2}. \end{aligned}$$

We now consider individual terms. It is assumed that the functions $A(z)$ and $B(z)$ have power series expansions valid for all argument values under consideration. Thus,

$$A(z)=\sum_{n=0}^{\infty} A_n z^n, \qquad B(z)=\sum_{n=0}^{\infty} B_n z^n,$$

$$x(\varepsilon)^* A^T(-j\varepsilon^r)A(j\varepsilon^r)x(\varepsilon) \leqq \|A(j\varepsilon^r)\|_2^2 \leqq \rho^2,$$

$$\begin{aligned} x(\varepsilon)^*[B^T(-j\varepsilon^{1-r})A(j\varepsilon^r)+A(-j\varepsilon^r)B(j\varepsilon^{1-r})]x(\varepsilon) \\ &= x(\varepsilon)^*\{B_1^T\cdot(-j\varepsilon^{1-r})\cdot A_0 + A_0^T\cdot B_1\cdot(j\varepsilon^{1-r})+O(\varepsilon)\}x(\varepsilon) \\ &= x(\varepsilon)^*\{[A_0^T B_1 - B_1^T A_0]\cdot j\varepsilon^{1-r}+O(\varepsilon)\}x(\varepsilon) \\ &= O(\varepsilon) \qquad (\varepsilon\to 0). \end{aligned}$$

This follows because the matrix in brackets is skew-symmetric, giving a zero quadratic form value.

Clearly, the remaining term is also $O(\varepsilon)$ because of the restricted range of r. Thus,

$$x(\varepsilon)^* B^T(-j\varepsilon^{1-r})B(j\varepsilon^{1-r})x(\varepsilon)=O(\varepsilon) \qquad (\varepsilon\to 0).$$

We now have

$$\begin{aligned} \|A(j\varepsilon^r)+B(j\varepsilon^{1-r})\|_2 &= \{\|A(j\varepsilon^r)x(\varepsilon)\|_2^2+O(\varepsilon)\}^{1/2} \\ &= \|A(j\varepsilon^r)x(\varepsilon)\|_2\cdot\{1+O(\varepsilon)\}^{1/2} \\ &\leqq \rho+O(\varepsilon). \end{aligned}$$

The middle equality above follows because $\|A(j\varepsilon^r)x(\varepsilon)\|_2$ is bounded away from zero. The last statement in the lemma is easily verified. □

We are now ready to prove the unproved portion of Theorem 3.1:

$$\begin{aligned} \|\hat{S}(j\omega,\varepsilon)\|_\infty &= \sup_{\omega\in R}\|S_S(j\omega)+S_F(j\varepsilon\omega)-S_S(\infty)\|_2, \\ &= \max\left\{\sup_{\omega\in[0,1/\sqrt{\varepsilon}]}\|S_S(j\omega)+S_F(j\varepsilon\omega)-S_S(\infty)\|_2,\right. \\ &\qquad \left.\sup_{\omega\in[1/\sqrt{\varepsilon},\infty)}\|S_S(j\omega)+S_F(j\varepsilon\omega)-S_S(\infty)\|_2\right\}. \end{aligned} \tag{A6}$$

We will show that the first term in braces satisfies the required inequality. The proof for the second term is done by an analogous process. Observe that for any $\varepsilon>0$ and $R>1$, the following set equality is valid:

$$[0,1/\sqrt{\varepsilon}]=[0,R]\cup\{\varepsilon^{-r}: 0<r\leqq\tfrac{1}{2}\}.$$

The function $S_S(s)$ has a power series expansion at ∞ that is valid for all $|s|>R_1$, for some R_1:

$$\begin{aligned} S_S(s) &= S_S(\infty)+\frac{S_{S1}}{s}+\frac{S_{S2}}{s^2}+\cdots \quad |s|>R_1. \\ &\triangleq P\left(\frac{1}{s}\right). \end{aligned} \tag{A7}$$

Let $R=\max\{1, R_1\}$.

The first term in braces in (A6) can be rewritten as

$$\sup_{\omega\in[0,1/\sqrt{\varepsilon}]}\|\hat{S}(j\omega,\varepsilon)\|_2 = \max\left\{\sup_{\omega\in[0,R]}\|\hat{S}(j\omega,\varepsilon)\|_2,\ \sup_{r\in(0^+,1/2]}\|\hat{S}(j\varepsilon^{-r},\varepsilon)\|_2\right\} \tag{A8}$$

The symbol 0^+ means that we do not sup over any r for which $\varepsilon^{-r} < R$. The first term in braces in (A8) is easily seen to be bounded by $\rho + O(\varepsilon)$. The second term in braces in (A8) can be rewritten, using (A7), as

$$\sup_{r\in(0^+,1/2]} \|\hat{S}(j\varepsilon^{-r}, \varepsilon)\|_2 = \sup_{r\in(0^+,1/2]} \|S_S(j\varepsilon^{-r}) + S_F(j\varepsilon^{1-r}) - S_S(\infty)\|_2$$

$$= \sup_{r\in(0^+,1/2]} \|P(-j\varepsilon^{r}) + S_F(j\varepsilon^{1-r}) - S_S(\infty)\|_2.$$

The result follows after making the identifications

$$A(z) = P(-z), \qquad B(z) = S_F(z) - S_S(\infty),$$

and applying Lemma A1. One final remark is in order. The $O(\varepsilon)$ result seems to require a lengthy proof. The weaker result

$$\|\hat{S}(j\omega, \varepsilon)\|_\infty \leqq \rho + O(\sqrt{\varepsilon})$$

is an almost immediate consequence of (A6).

REFERENCES

[1] A. G. J. MacFarlane, ed., *Frequency Response Methods in Control Systems*, IEEE Press, New York, 1979.

[2] G. Zames, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301–320.

[3] M. Vidyasagar, *Control Systems Synthesis: A Factorization Approach*, The MIT Press, Cambridge, MA, 1985.

[4] K. Hoffman, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1965.

[5] J. W. Helton, *Worst case analysis in the frequency domain: the H^∞ approach to control*, IEEE Trans. Automat. Control, 30 (1985), pp. 1154–1170.

[6] J. A. Ball and J. W. Helton, *A Beurling-Lax Theorem for the Lie group $U(m, n)$ which contains most classical interpolation theory*, J. Operator Theory, 9 (1983), pp. 107–142.

[7] D. W. Luse and H. K. Khalil, *Frequency domain results for systems with slow and fast dynamics*, IEEE Trans. Automat. Control, 30 (1985), pp. 1171–1179.

[8] D. W. Luse, *Frequency domain results for systems with multiple time scales*, IEEE Trans. Automat. Control, 31 (1986), pp. 918–924.

[9] L. V. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1979.

[10] D. W. Luse, *State-space realization of multiple-frequency-scale transfer matrices*, IEEE Trans. Automat. Control, 33 (1988), pp. 185–187.

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

[12] J. W. Brewer, J. W. Bunce, and F. S. Van Vleck, *Linear Systems Over Commutative Rings*, Marcel Dekker, New York, 1986.

[13] V. R. Saksena, J. O'Reilly, and P. V. Kokotovic, *Singular perturbation and time scale methods in control theory: survey* 1976–1983, Automatica, 20 (1984), pp. 273–293.

[14] H. K. Khalil, *Output feedback control of linear two-time-scale systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 784–792.

[15] J. H. Chow and P. V. Kokotovic, *A decomposition of near-optimum regulators for systems with slow and fast modes*, IEEE Trans. Automat. Control, 21 (1976), pp. 701–705.

[16] J. A. Ball and A. C. M. Ran, *Global inverse spectral problems for rational matrix functions*, Linear Algebra Appl., 86 (1987), pp. 273–282.

[17] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[18] H. K. Khalil and Z. Gajic, *Near-optimum regulators for stochastic linear singularity perturbed systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 531–541.

[19] H. Bart, I. Gohberg, and M. A. Kaashoek, *Minimal Factorization of Matrix and Operator Functions*, OT1, Birkhauser, Basel, 1979.

[20] J. A. Ball and D. W. Luse, *Sensitivity minimization as a Nevanlinna-Pick interpolation problem*, Proc. of the NATO Advanced Research Workshop on Modelling Robustness, and Sensitivity Reduction in Control Systems, Groningen, the Netherlands, December 1–5, 1986.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 836-841, July 1989

009

COMPLEMENTARITY PROBLEMS OVER LOCALLY COMPACT CONES*

M. SEETHARAMA GOWDA†

Abstract. This paper proves an existence result for a complementarity problem (on a locally convex space) where the mapping is copositive, positive homogeneous, and of monotone type on a locally compact cone. A perturbation theorem is proved that extends a result of Mangasarian and Doverspike proved for $n \times n$ matrices on the nonnegative orthant.

Key words. copositive, positive homogeneous, locally compact cone, complementarity problem, Mackey topology

AMS(MOS) subject classifications. 49A99, 90C33

1. Introduction. Let (X, τ) be a real locally convex space, let K be a closed convex cone in X, and let q be in X^*. For a mapping $T: K \to X^*$, the *Complementarity Problem* (denoted by CP(X, T, K, q) or by CP(T, K, q) when X is fixed) is to find

$$x \in K \quad \text{such that } Tx + q \in K^* \quad \text{and} \quad \langle Tx + q, x \rangle = 0.$$

Here $K^* := \{y \in X^* : \langle y, k \rangle \geqq 0 \text{ for all } k \in K\}$, where $\langle y, x \rangle = y(x)$ for $y \in X^*$, $x \in X$. In the finite-dimensional setting, (linear) complementarity problems are related to linear (and quadratic) programming, bimatrix games, etc. (see Cottle and Dantzig [3]). Certain problems in engineering and economics can be posed as (linear) complementarity problems. In the infinite-dimensional setting, complementarity problems are related to variational inequalities. They also appear in certain engineering problems (see Cryer and Dempster [4], Isac [11], and references therein).

In this article, we prove an existence result (Theorem 1) for a copositive T that is positive homogeneous and of monotone type on a locally compact cone. We also state a perturbation result (Theorem 2) that extends a result of Mangasarian and Doverspike [16], [5] stated for matrices on $\mathbb{R}^n$. Banach space applications of our results can be obtained by working with the weak topology. Our results are new even in the finite-dimensional setting and generalize our earlier results in [8].

2. Preliminaries. (X, τ) denotes a real locally convex space. K, X^*, q, and T, respectively, denote a closed convex cone in X, the dual of X, an element of X^*, and a mapping from K into X^*. $\sigma(X, X^*)$ stands for the weak topology on X and $\sigma(X^*, X)$ denotes the *weak** topology on X^*. $\mathcal{M}(X^*, X)$ denotes the Mackey topology on X^*, which is the topology of uniform convergence on balanced, convex, $\sigma(X, X^*)$-compact subsets of X. (When X is a reflexive Banach space, Mackey topology coincides with the norm topology.) We use the result $(X^*, \mathcal{M}(X^*, X))^* = X$ (see Horvath [9, p. 205]). For $f \in X^*$, $x \in X$, $\langle f, x \rangle$ denotes $f(x)$. For $E \subseteq X$, $F \subseteq X^*$, we define $E^* := \{f \in X^* : \langle f, x \rangle \geqq 0 \text{ for all } x \in E\}$, $F^* := \{x \in X : \langle f, x \rangle \geqq 0 \text{ for all } f \in F\}$. We denote the *solution set* of CP(T, K, q) by Sol(T, K, q). We note that (since K is a cone) $x \in \text{Sol}(T, K, q)$ if and only if $\langle Tx + q, y - z \rangle \geqq 0$ for all $y \in K$. We say that:

(a) T is *copositive* on K if $\langle Tx, x \rangle \geqq 0$ for all $x \in K$.

(b) T is *copositive plus* on K if T is copositive on K and $x \in K$, $\langle Tx, x \rangle = 0 \Rightarrow \langle Tx, k \rangle + \langle Tk, x \rangle = 0$ for all $k \in K$.

* Received by the editors April 25, 1988; accepted for publication (in revised form) October 5, 1988.

† Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, Maryland 21228.

(c) T is *copositive star* on K if T is copositive on K and $x \in K$, $Tx \in K^*$, $\langle Tx, x\rangle = 0 \Rightarrow -x \in (T(K))^*$.

(d) T is *pseudomonotone* on K if for all $x, y \in K$, $\langle Ty, x-y\rangle \geqq 0 \Rightarrow \langle Tx, x-y\rangle \geqq 0$.

(e) T is *positive homogeneous on* K if there is a positive γ such that $T(\lambda x) = \lambda^{\gamma} Tx$ for all $\lambda \geqq 0$, for all $x \in K$.

(f) T is of *monotone type* on K if either

(i) $x \to \langle Tx, x-y\rangle$ is lower semicontinuous on (K, τ) for each fixed $y \in K$, or

(ii) T is monotone on K (i.e., $\langle Tx - Ty, x-y\rangle \geqq 0$, for all $x, y \in K$) and semicontinuous on K (i.e., continuous from lines in K to $\sigma(X^*, X)$).

It can be easily shown that copositive plus mappings as well as pseudomonotone, positive homogeneous mappings are copositive star. In this article we deal with locally compact cones. Examples of locally compact cones include finite-dimensional cones and cones in the dual X^* of a normed linear space defined, for any $\alpha \in (0, 1)$ and $e \in X$, by $K = \{f \in X^*: \alpha \|f\| \leqq f(e)\}$.

When K is locally compact, K can be written as the direct sum of a finite dimensional subspace $M (= K \cap -K)$ of X and a cone L with compact base B given by $B = \{x \in L: \langle e, x\rangle = 1\}$ for some $e \in K^*$. (This follows from Thm. 3.12.8 of [13].)

For $x \in K$ we can write $x = \lambda b + m$ ($\lambda \geqq 0$, $m \in M$, $b \in B$) and define $g(x) := \lambda + \|m\| (= \langle e, x\rangle + \|Pm\|$ where P is the projection of X onto M). We see that $g(x)$ is continuous on (K, τ) and that the set $\{x \in K: g(x) = 1\}$ is compact. We observe that $\{x \in K: g(x) \leqq 1\}$ is compact and convex.

3. Results. Throughout this section we assume the following:

(a) T is copositive, positive homogeneous, and of monotone type on K;

(b) K is locally compact in (X, τ).

Let $S := \{x \in K: Tx \in K^*, \langle Tx, x\rangle = 0\}$. Recall that Sol (T, K, q) is the set of all solutions of CP (T, K, q).

THEOREM 1. *Suppose that* $0 \neq x \in S$ *implies* $\langle q, x\rangle > 0$. *Then* Sol $(T, K, q) \neq \emptyset$. *If further* $T: (K, \tau) \to (X^*, \sigma(X^*, X))$ *is continuous, then* Sol (T, K, q) *is compact in* (X, τ).

Proof. For the function g defined in § 2, we see that the set $\{x \in K: g(x) \leqq 1, \langle q, x\rangle \leqq 0\}$ is compact and convex. The argument in Theorem 1 (ii) of Borwein [2] then shows that the variational inequality

$$\tilde{H}(g, g, q): \langle Tx + (1 - g(x))q, y - x\rangle \geqq 0 \quad \forall y \in K, \quad g(y) \leqq 1, \quad \langle q, y\rangle \leqq 0$$

is solvable for some $x \in K$ with $g(x) \leqq 1$ and $\langle q, x\rangle \leqq 0$. Fix this x and define $f(y) := \langle Tx + (1 - g(x))q, y - x\rangle$, $(y \in K)$. Since zero solves CP (T, K, q) when $q \in K^*$, we can assume that $q \notin K^*$. Then there is a $k \in K$ such that $g(k) < 1$ and $\langle q, k\rangle < 0$. Thus, the convex program

$$f(x) = \min \{f(y): g(y) \leqq 1, \langle q, y\rangle \leqq 0, y \in K\}$$

satisfies the Slater condition. Hence, there are $t \geqq 0$ and $s \geqq 0$ such that

$$f(x) \leqq f(y) + t(g(y) - 1) + s\langle q, y\rangle \quad (\forall y \in K) \tag{3.1}$$

and $t(g(x) - 1) = 0 = s\langle q, x\rangle$ (cf. [10, Thm. 2, p. 68]). If $g(x) < 1$, then $t = 0$, and (3.1) leads to

$$\langle Tx + (1 - g(x) + s)q, y - x\rangle \geqq 0$$

(since $f(x) = 0$ and $s\langle q, x\rangle = 0$). Also, $g(x) < 1$ gives $1 - g(x) + s > 0$; hence (by homogeneity of T), $(1 - g(x) + s)^{-1/\gamma} x$ solves CP (T, K, q).

Now suppose that $g(x)=1$. Writing (3.1) in the subdifferential form (see, e.g., [10, Thm. 2′, p. 69]), we get

$$0 \in \partial f(x) + t\partial g(x) + sq - (K-x)^*.$$

In this case,

$$\langle Tx + tp + sq, y - x\rangle \geqq 0 \quad (\forall y \in K) \tag{3.2}$$

for some $p \in \partial g(x)$. From this we get $\langle Tx + tp + sq, x\rangle = 0$, i.e., $\langle Tx, x\rangle + t\langle p, x\rangle = 0$. Since $p \in \partial g(x)$ we have $\langle p, x\rangle \geqq g(x) = 1$; hence, $\langle Tx, x\rangle = 0$ (by copositivity of T) and $t\langle p, x\rangle = 0$, which leads to $t=0$. If $s=0$, then (3.2) gives $Tx \in K^*$ and $\langle Tx, x\rangle = 0$, i.e., $x \in S$. Since $g(x)=1$ and $\langle q, x\rangle \leqq 0$, this leads to a contradiction. Hence $s>0$, and in this case (3.2) shows (since $t=0$ and T is positive homogeneous) that $s^{-(1/\gamma)}x$ solves CP(T, K, q).

Now suppose that $T: (K, \tau) \to (X^*, \sigma(X^*, X))$ is continuous. From this it follows that the solution set of CP(T, K, q) is closed. To get compactness of this set (from the local compactness of K) it is enough to show that the solution set is bounded. Suppose, if possible, that $\{x_\alpha\}$ is an unbounded net (consisting of nonzero elements) in the solution set. Using the decomposition $K = L \oplus M$, we can write (as in § 2), $x_\alpha = \lambda_\alpha b_\alpha + m_\alpha$ and observe from the nonnegativity of λ_α that $g(x_\alpha) = \lambda_\alpha + \|m_\alpha\|$ is an unbounded net in $\mathbb{R}_+$. The net $\{g(x_\alpha)^{-1}x_\alpha\}$ is contained in the compact set $\{x \in K: g(x) = 1\}$; hence, a subnet $\{g(x_\beta)^{-1}x_\beta\}$ converges to, say, $\bar{x}$ (in K) such that $g(\bar{x})=1$.

Then $\langle Tx_\beta + q, y\rangle \geqq 0$ (for all $y \in K$, for all β) leads to $\langle T(g(x_\beta)^{-1}x_\beta) + qg(x_\beta)^{-\gamma}, y\rangle \geqq 0$, which, upon taking limits (and using the continuity of T), gives

$$\langle T\bar{x}, y\rangle \geqq 0 \quad (\forall y \in K). \tag{3.3}$$

Also,

$$\langle Tx_\beta, x_\beta\rangle + \langle q, x_\beta\rangle = 0 \quad (\forall \beta) \tag{3.4}$$

leads to $\langle q, x_\beta\rangle \leqq 0$ (by copositivity) and hence to $\langle q, g(x_\beta)^{-1}x_\beta\rangle \leqq 0$. Upon taking limits we get

$$\langle q, \bar{x}\rangle \leqq 0. \tag{3.5}$$

Further, (3.4) gives

$$\langle Tg(x_\beta)^{-1}x_\beta, g(x_\beta)^{-1}x_\beta\rangle + g(x_\beta)^{-\gamma}\langle q, g(x_\beta)^{-1}x_\beta\rangle = 0 \quad (\forall \beta).$$

Now continuity of T gives

$$\langle T\bar{x}, \bar{x}\rangle = 0. \tag{3.6}$$

We see that $\bar{x} \in S$ and $\langle q, \bar{x}\rangle \leqq 0$. This contradicts our hypothesis, since $g(\bar{x}) = 1$ implies that $\bar{x} \neq 0$. Thus Sol(T, K, q) is bounded, and compactness of this set follows. □

Remark 1. When $X = \mathbb{R}^n$ and T is an $n \times n$ matrix copositive on $\mathbb{R}^n_+$, LCP$(T, \mathbb{R}^n_+, q)$ is solvable for all $q \in S^*$. This is a result of Lemke [15, p. 104]. It can be easily shown that this same result holds when $\mathbb{R}^n_+$ is replaced by a polyhedral cone (Gowda [6]). However, for general cones, the implication $0 \neq x \in S \Rightarrow \langle q, x\rangle \geqq 0$ need not give the solvability of CP(T, K, q). This can be easily seen by taking in $\mathbb{R}^3$, $K = \{(x, y, z): x, z \geqq 0, 2xz \geqq y^2\}$, $T(x, y, z) = (x, y, 0)$, and $q = (1, 1, 0)$. (We observe that in this example, T is positive semidefinite and that CP(T, K, q) is feasible. Thus, on a nonpolyhedral cone, feasibility need not give solvability even for a positive semidefinite matrix.)

THEOREM 2. *Consider the following*:

(a) q *belongs to the* $\mathcal{M}$*-interior of the closed convex cone generated by* $K^* - T(K)$.

(b) $0 \neq x \in S \Rightarrow \langle q, x\rangle > 0$.

(c) CP(T, K, q) *has a solution.*

(d) CP(T, K, q) has a nonempty compact solution set.

We have the following:

(i) *If* T *is copositive star on* K *then* (a)$\Rightarrow$(b)$\Rightarrow$(c).

(ii) *If* T *is copositive star and* T *is continuous from* (K, τ) *into* $(X^*, \sigma(X^*, X))$, *then* (a)$\Rightarrow$(b)$\Rightarrow$(d).

(iii) *If* T *is copositive plus, linear, and continuous from* (K, τ) *into* $(X^*, \sigma(X^*, X))$, *then* (a)$\Leftrightarrow$(b)$\Leftrightarrow$(d).

Proof. (i) Let T be copositive star on K. In view of Theorem 1, we need only show that (a)$\Rightarrow$(b) and to this end suppose that (b) is false. Then there exist $x \neq 0$, $x \in S$, and $\langle q, x\rangle \leqq 0$. We have $-x \in (TK)^*$; hence,

$$\langle q, x\rangle \leqq 0 \leqq \langle k^* - T(k), x\rangle \quad \forall k^* \in K^*, \quad k \in K.$$

This says that q can be separated from $K^* - T(K)$ by a hyperplane, i.e., (a) is false.

(ii) This part follows from Theorem 1 and (i).

(iii) (d)$\Rightarrow$(b): Suppose there is a $d(\neq 0)$ in S such that $\langle q, d\rangle \leqq 0$. Let x be any solution of CP(T, K, q). Since T is copositive plus, we have $\langle Tx, d\rangle + \langle Td, x\rangle = 0$. We see from the linearity of T that $T(x + \lambda d) + q \in K^*$ (for all $\lambda \geqq 0$) and

$$\begin{aligned}\langle T(x+\lambda d), x+\lambda d\rangle &= \langle Tx + q, x\rangle + \lambda\{\langle Tx, d\rangle + \langle Td, x\rangle\} + \lambda^2\langle Td, d\rangle + \langle q, d\rangle \\ &= \langle q, d\rangle \leqq 0.\end{aligned}$$

But this says that $x + \lambda d$ solves CP(T, K, q) for all $\lambda \geqq 0$, contradicting the compactness of the solution set. Thus (d)$\Rightarrow$(b).

(b)$\Rightarrow$(a): Suppose that (a) is false. Then there is a net $\{q_\alpha\}$ such that $q_\alpha \notin \mathcal{M}$-closed convex cone generated by $K^* - T(K)$ and $q_\alpha \to q$ in $(X^*, \mathcal{M})$. Using $(X^*, \mathcal{M})^* = X$ and a separation theorem, we get $x_\alpha (\neq 0) \in X$ such that

$$\langle q_\alpha, x_\alpha\rangle < 0 \leqq \langle k^* - T(k), x_\alpha\rangle \quad \forall k^* \in K^*, \quad k \in K, \quad \forall \alpha.$$

This shows that $x_\alpha \in K$, $\langle Tx_\alpha, x_\alpha\rangle = 0$, and $\langle -Tk, x_\alpha\rangle \geqq 0$ for all $k \in K$, for all α. From this we get

$$\begin{aligned}\langle Tx_\alpha, k\rangle &= \langle Tx_\alpha, k\rangle + \langle Tk, x_\alpha\rangle - \langle Tk, x_\alpha\rangle \\ &= 0 - \langle Tk, x_\alpha\rangle \quad \text{(since } T \text{ is copositive plus)} \\ &\geqq 0 \quad \forall k \in K, \quad \forall \alpha.\end{aligned}$$

Thus $0 \neq x_\alpha \in S$. As in the proof of Theorem 1, we can show that $g(x_\alpha)^{-1}x_\alpha$ has a subnet $g(x_\beta)^{-1}x_\beta$ converging to, say, $\bar{x}$. This $\bar{x} \neq 0$ and is in S. Now the set $E := \{x \in K : g(x) \leqq 1\}$ is compact convex in τ and hence in $\sigma(X, X^*)$. The balanced hull $[-1, 1]E$ of this set is also compact convex in $\sigma(X, X^*)$.

Since $q_\beta \to q$ in $\mathcal{M}$ and $g(x_\beta)^{-1}x_\beta \to \bar{x}$ in E, we have

$$\langle q, \bar{x}\rangle = \lim \langle q_\beta, g(x_\beta)^{-1}x_\beta\rangle \leqq 0.$$

Thus $0 \neq \bar{x} \in S$ and $\langle q, \bar{x}\rangle \leqq 0$, i.e., (b) fails.

Hence (b)$\Rightarrow$(a). □

Remark 2. The proof of (b)$\Rightarrow$(a) in (iii) is valid if T is copositive, linear, and continuous (because in this situation, $x \in K$, $\langle Tx, x\rangle = 0 \Rightarrow \langle Tx, k\rangle + \langle Tk, x\rangle \geqq 0$ for all $k \in K$).

Remark 3. Theorem 2(iii) generalizes our earlier Hilbert space results (Theorems 4.1 and 6.1 of [8]), where a condition equivalent to the local compactness condition has been used. When $X=\mathbb{R}^n$, $K=\mathbb{R}^n_+$, and T is an $n\times n$ copositive plus matrix, the equivalence of (a) and (d) is the well-known result of Mangasarian and Doverspike [16], [5] that CP$(T,\mathbb{R}^n_+,\bar{q})$ is feasible for all $\bar{q}$ near q and only if CP$(T,\mathbb{R}^n_+,q)$ has compact solution set.

Remark 4. In [1] Allen proves that if there is an $x_0\in K$ such that $F(x_0)\in \mathcal{M}$-int(K^*), where $F(x):=Tx+q$ is pseudomonotone and $\langle Tx, x\rangle$ is weak lower semicontinuous on K, then CP(T,K,q) has a solution. (As Isac [12] shows, we need the weak lower semicontinuity of $x\to\langle Tx, x-y\rangle$ for any $y\in K$ to get this result.) This is different from our results even for matrices. While Allen assumes the pseudomonotonicity of $x\to Tx+q$ and gets the solvability of CP(T,K,q), we assume the pseudomonotonicity (rather, the copositive star property) of T and get the solvability of CP(T,K,q) for *any* $q\in$ int$(K^*-T(K))$. We wish to note that pseudomonotonicity of T need not imply the pseudomonotonicity of the mapping $x\to Tx+q$. For example, let

$$T=\begin{bmatrix}0 & -1\\ 2 & 0\end{bmatrix},\quad q=\begin{bmatrix}1\\ -2\end{bmatrix},\quad u=\begin{bmatrix}1\\ 1\end{bmatrix},\quad v=\begin{bmatrix}2\\ 0\end{bmatrix}.$$

It is easily seen that T is pseudomonotone on $\mathbb{R}^2_+$ (see, e.g., [7]) while the mapping $x\to Tx+q$ is not (since $\langle Tu+q, v-u\rangle=0$ and $\langle Tv+q, v-u\rangle=-1$).

THEOREM 3. *Consider the following*:

(a) $S=\{0\}$;

(b) CP(T,K,q) *has a solution for all* $q\in X^*$;

(c) CP(T,K,q) *is feasible for all* $q\in X^*$.

We have the following:

(i) (a)$\Rightarrow$(b)$\Rightarrow$(c).

(ii) *If* T *is copositive star on* K *then* $(a)\Leftrightarrow(b)\Leftrightarrow(c)$.

Proof. (i) (a)$\Rightarrow$(b) follows from Theorem 1. (b)$\Rightarrow$(c) is obvious.

(ii) Let T be copositive star on K. If (a) is false then there is a $d\neq 0$ such that $d\in K$, $Td\in K^*$, and $\langle Td, d\rangle=0$. Since T is copositive star we have $\langle Tx, d\rangle\leqq 0$, for all $x\in K$. Hence $\langle Tx-d, d\rangle<0$, showing the infeasibility of CP$(T,K,-d)$. Thus (c) is false and (c)$\Rightarrow$(a). □

Remark 5. Theorem 3(i) also follows from Theorem 5 in Borwein [2]. Theorem 3(ii) extends Corollary 7 in Borwein [2], where it is proved for copositive plus (linear) mappings.

REFERENCES

[1] G. ALLEN, *Variational inequalities, complementarity problems, and duality theorems*, J. Math. Anal. Appl., 58 (1977), pp. 1-10.

[2] J. M. BORWEIN, *Alternative Theorems and General Complementarity Problems*, Lecture Notes in Economics and Mathematical Systems 259, Springer-Verlag, Berlin, New York, 1985, pp. 194-203.

[3] R. W. COTTLE AND G. B. DANTZIG, *Complementary pivot theory of mathematical programming*, in Mathematics of Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., eds., American Mathematical Society, Providence, RI, 1968.

[4] C. W. CRYER AND M. A. H. DEMPSTER, *Equivalence of linear programs in vector lattice Hilbert spaces*, SIAM J. Control Optim., 18 (1980), pp. 76-90.

[5] R. DOVERSPIKE, *Some perturbation results for the linear complementarity Problem*, Math. Programming, 23 (1982), pp. 181-192.

[6] M. S. GOWDA, *Linear complementarity problems*, Research Rept., University of Maryland, Baltimore County, Baltimore, MD, September 1987.

[7] M. S. GOWDA, *Pseudomonotone and copositive star matrices*, Research Rept., University of Maryland, Baltimore County, Baltimore, MD, March 1987; Linear Algebra Appl., to appear.

[8] M. S. GOWDA AND T. I. SEIDMAN, *Generalized linear complementarity problems*, Research Report, University of Maryland, Baltimore County, Baltimore, MD, August 1985; Math. Programming, to appear.

[9] J. HORVATH, *Topological Vector Spaces and Distributions*, Vol. 1, Addison-Wesley, Reading, MA, 1966.

[10] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, North-Holland, New York, 1979.

[11] G. ISAC, *Nonlinear complementarity problem and Galerkin method*, J. Math. Anal. Appl., 108 (1985), pp. 563-574.

[12] ———, *On some generalization of Karamardian's theorem on the complementarity problem*, Preprint, Dept. Collège Militaire Royal de Saint-Jean, Saint-Jean, Québec, Canada, 1987.

[13] G. JAMESON, *Ordered Linear Spaces*, Lecture Notes in Math. 141, Springer-Verlag, Berlin, New York, 1970.

[14] S. KARAMARDIAN, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl., 18 (1976), pp. 445-454.

[15] C. E. LEMKE, *On complementary pivot theory*, in Mathematics of Decision Sciences, Part I, G. B. Dantzig and A. F. Veinott, Jr., ed., American Mathematical Society, Providence, RI, 1968.

[16] O. L. MANGASARIAN, *Characterizations of bounded solutions of linear complementarity problems*, Math. Programming Stud., 19 (1982), pp. 153-166.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 842-860, July 1989

010

WEIGHTED OPTIMIZATION THEORY FOR NONLINEAR SYSTEMS*

CIPRIAN FOIAS† AND ALLEN TANNENBAUM‡

Abstract. In this paper, the solution of a nonlinear version of the weighted sensitivity H^∞-optimization problem is discussed. It is shown that the natural object to be considered in this context is a certain "sensitivity operator," which will be optimized locally in a given "energy ball" (see § 5 for the details). In the linear case, the authors are reduced again to the classical sensitivity minimization technique of Zames [21]. The methods were very strongly influenced by the complex analytic power series ideas of [3], [4], [5]. See also the recent results of Ball and Helton [6] for another approach to this subject.

Key words. sensitivity operator, nonlinear control, dilation theory, skew Toeplitz operator

AMS(MOS) subject classifications. 93B35, 93C05

1. Introduction. Recently, there has been a great deal of research devoted to the weighted H^∞-optimization of linear systems. See [13] for a rather extensive list of references. Much of the underlying theory for this work has been based on the ideas of Adamjan, Arov, and Krein [1], generalized interpolation theory in H^∞ due to Sarason [17], and, most generally, on the Sz.-Nagy–Foias commutant lifting theorem [19].

In the papers [3], [4] an extension of the commutant lifting theorem to a local nonlinear setting was given, together with a discussion of how this result could be used to develop a design procedure for nonlinear systems. In the present paper, we continue this line of research with a constructive extension of the linear H^∞ theory to nonlinear systems. We should note that our colleagues Ball and Helton [6] have developed a completely different, novel approach to this problem based on a nonlinear version of Ball–Helton theory.

In the theory presented below, we will consider majorizable input/output operators (see § 3 for the precise definition). In particular, these operators are analytic in a ball around the origin in a complex Hilbert space, and it turns out that it is possible to express each n-linear term of the Taylor expansion of such an operator as a linear operator on a certain tensor space. (Our class of operators also includes Volterra series of fading memory [8].) This allows us to iteratively apply the classical commutant lifting theorem in designing a compensator. (The general technique we call the *iterative commutant lifting procedure.* See § 6 for the details.) For single input/single output (SISO) systems, this leads to the construction of a compensator which is optimal relative to a certain sensitivity function that will be defined in § 5. Moreover, in complete generality (i.e., for multiple input/multiple output (MIMO) systems), our procedure will ameliorate (in the sense of our nonlinear weighted sensitivity criterion) any given design. We note that for linear systems, our method reduces to the standard H^∞ design technique as discussed, for example, in [13] and initiated in [21].

In developing the present theory, we have had to extend some of the skew Toeplitz techniques of [7] and [11] to linear operators defined on certain tensor spaces. This

* Received by the editors May 16, 1988; accepted for publication (in revised form) November 20, 1988. This research was supported in part by grants from the Research Fund of Indiana University, the Department of Energy (DE-FG02-86ER25020), the National Science Foundation (ECS-8704047) and (DMS-8811084), and the Air Force Office of Scientific Research (AFOSR-88-0020).

† Department of Mathematics, Indiana University, Bloomington, Indiana 47405.

‡ Department of Electrical Engineering, University of Minnesota, 123 Church Street SE, Minneapolis, Minnesota 55455.

has led to several novel results in computational operator theory, and, for example, provides a way of iteratively constructing the nonlinear intertwining dilation of the nonlinear commutant lifting theorem considered in [3] and [4]. Moreover, we provide a generalization of a formula due to Sarason [17] for the optimal interpolant in terms of a maximal vector. See § 8 for the details.

An important point is that many of our results are constructive and lead to physically implementable compensators. In fact, we reduce a nonlinear optimization problem to an iterative linear procedure, each step of which we know how to solve. This is illustrated by an example in § 9.

2. Analytic mappings on Hilbert space. We would like to discuss here a few standard results about analytic mappings on Hilbert spaces. We are essentially following the treatments of [3]–[5] and [8] to which the reader may refer for all of the details. In particular, input/output operators that admit Volterra expansions are special cases of the operators which we study here. See [8], [16], [20].

Let G and H denote complex Hilbert spaces. Set

$$B_{r_o}(G) := \{g \in G: \|g\| < r_o\}$$

(the open ball of radius r_o in G about the origin). Then we say that a mapping $\phi: B_{r_o}(G) \to H$ is *analytic* if the complex function $(z_1, \cdots, z_n) \to \langle \phi(z_1 g_1 + \cdots + z_n g_n), h\rangle$ is analytic in a neighborhood of $(1, 1, \cdots, 1) \in \mathbf{C}^n$ as a function of the complex variables $z_1, \cdots, z_n$ for all $g_1, \cdots, g_n \in G$ such that $\|g_1 + \cdots + g_n\| < r_o$, for all $h \in H$, and for all $n > 0$. (Note that we denote the Hilbert space norms in G and H by $\|\ \|$ and the inner products by $\langle\ ,\ \rangle$.)

We will now assume that $\phi(0) = 0$. It is easy to see that if $\phi: B_{r_o}(G) \to H$ is analytic, then ϕ admits a convergent Taylor series expansion, i.e.,

$$\phi(g) = \phi_1(g) + \phi_2(g, g) + \cdots + \phi_n(g, \cdots, g) + \cdots,$$

where $\phi_n: G \times \cdots \times G \to H$ is an n-linear map. Clearly, without loss of generality we may assume that the n-linear map $(g_1, \cdots, g_n) \to \phi_n(g_1, \cdots, g_n)$ is symmetric in the arguments $g_1, \cdots, g_n$. This assumption will be made throughout this paper for the various analytic maps that we consider. For ϕ a Volterra series, ϕ_n is basically the nth-Volterra kernel.

Now set

$$\hat{\phi}_n(g_1 \otimes \cdots \otimes g_n) := \phi_n(g_1, \cdots, g_n).$$

Then $\hat{\phi}_n$ extends in a unique manner to a dense subset of $G^{\otimes n} := G \otimes \cdots \otimes G$ (tensor product taken n times). Note by $G^{\otimes n}$ we mean the Hilbert space completion of the algebraic tensor product of the G's. Clearly if $\hat{\phi}_n$ has finite norm on this dense subset, then $\hat{\phi}_n$ extends by continuity to a bounded linear operator $\hat{\phi}_n: G^{\otimes n} \to H$. By abuse of notation, we will set $\phi_n := \hat{\phi}_n$, and $\phi_n(g) := \phi_n(g \otimes \cdots \otimes g)$ (the tensor product taken n times).

It is important to note that in principle we can determine ϕ_n quite easily from the input/output operator ϕ. Indeed, we have the following elementary lemma.

LEMMA 2.1. *Let $\phi: B_{r_o}(G) \to H$ be analytic, $\phi(0) = 0$. Suppose, moreover, that if*

$$\phi(g) = \phi_1(g) + \cdots \phi_n(g) + \cdots,$$

then each of the ϕ_n defines a bounded linear operator $G^{\otimes n} \to H$ as above (and is symmetric in its arguments). Then for $g_j \in G$ $(j = 1, \cdots, n)$ with $\|g_1\| = \cdots = \|g_n\| < r_o$, we have

$$n!\,\phi_n(g_1 \otimes \cdots \otimes g_n) = \frac{1}{(2\pi)^n} \int_0^{2\pi} \cdots \int_0^{2\pi} \phi(\exp(i\theta_1)g_1 + \cdots + \exp(i\theta_n)g_n) \times \exp(-i(\theta_1 + \cdots + \theta_n))\, d\theta_1 \cdots d\theta_n.$$

Proof. Expand $\phi(z_1g_1+\cdots+z_ng_n)$ in powers of $z_1, \cdots, z_n$. Then it is easy to see that the coefficient of $z_1 \cdots z_n$ is precisely $n!\phi_n(g_1\otimes\cdots\otimes g_n)$. The required result then follows immediately from the Cauchy formula. □

Remark 2.2. We should note that if ϕ is analytic, then each ϕ_n is continuous (as an n-multilinear map); hence, the associated linear map extends to the nth projective power of G. Lemma 2.1 is valid in this more general situation as well.

We now conclude this section with two key definitions.

DEFINITION 2.3. (i) Notation as above. By a *majorizing sequence* for the holomorphic map ϕ, we mean a sequence of positive numbers α_n $n = 1, 2, \cdots$ such that $\|\phi_n\| < \alpha_n$ for $n \geqq 1$. Suppose that $\rho := \limsup \alpha_n^{1/n} < \infty$. Then it is completely standard ([8]) that the Taylor series expansion of ϕ converges at least on the ball $B_r(G)$ of radius $r = 1/\rho$.

(ii) If ϕ admits a majorizing sequence as in (i), then we will say that ϕ is *majorizable.*

We will see in the next section that a very important class of input/output operators from systems and control theory are in point of fact majorizable.

3. Operators with fading memory. In this section, we will show that perhaps the most natural class of input/output operators from the systems standpoint are majorizable. Moreover, for this class of operators we will even derive an a priori majorizing sequence. We begin with the following key definition:

DEFINITION 3.1. An analytic map $\phi: B_{r_0}(G) \to H$, $\phi(0) = 0$ has *fading memory* if its nonlinear part $\phi - \phi'(0)$ admits a factorization

$$\phi - \phi'(0) = \hat{\phi} \circ W,$$

where $\hat{\phi}$ is an analytic map defined in some neighborhood of $0 \in G$, and W is a linear Hilbert-Schmidt operator. (In this case, we can assume that there exists an orthonormal basis of eigenvectors for W in G, $\{e_k\}$, $k = 1, 2, \cdots$ such that $We_k = \lambda_k e_k$ with

$$\|W\|_2^2 := \sum_{k=1}^{\infty} |\lambda_k|^2 < \infty.$$

$\|W\|_2$ is called the *Hilbert-Schmidt norm* of W.)

Remark 3.2. System-theoretically fading memory input/output operators have the property that any two input signals, which are close in the recent past but not necessarily close in the remote past, will yield present outputs which are close. For more details about this important class of operators, see [8].

For fading memory operators, we can construct an explicit majorizing sequence.

LEMMA 3.3. *Let $\phi: B_{r_0}(G) \to H$, $\phi(0) = 0$, have fading memory. Suppose, moreover, that if we write*

$$\phi - \phi'(0) = \hat{\phi} \circ W$$

as in (3.1), *then $\hat{\phi}: B_{r_1}(G) \to B_{r_2}(H)$. Then the sequence*

$$\alpha_1 := \|\phi'(0)\|$$

$$\alpha_n := \frac{r_2 e^n \|W\|_2^n}{r_1^n}$$

for $n \geqq 2$, is a majorizing sequence for ϕ.

Proof. For complete details see [4, Lemma (3.5)]. However, since we will need some estimates from the proof for Proposition (3.5) below, we will give an outline.

First, without loss of generality we may assume that W is positive. Since $\hat{\phi}: B_{r_1}(G) \to B_{r_2}(H)$, from (2.1) we obtain

$$\|\phi_n(g_1 \otimes \cdots \otimes g_n)\| \leqq \frac{1}{n!} r_2,$$

for $\|g_1\| = \cdots = \|g_n\| \leqq r_1/n$.

Now, since $\{e_{i_1} \otimes \cdots \otimes e_{i_n}: 1 \leqq i_1, \cdots, i_n\}$ is an orthonormal basis of $G^{\otimes n}$, we can write $g \in G^{\otimes n}$ as

$$\sum_{1 \leqq i_1, \cdots, i_n}^{\infty} \alpha_{i_1 \cdots i_n} e_{i_1} \otimes \cdots \otimes e_{i_n}$$

and

$$\|\hat{g}\|^2 = \sum_{1 \leqq i_1, \cdots, i_n}^{\infty} |\alpha_{i_1 \cdots i_n}|^2 < \infty.$$

Now, from the above we can easily compute (see [4] for the details) that

$$\begin{aligned}
&\left\| \phi_n \left(\sum_{1 \leqq i_1, \cdots, i_n}^{\infty} \alpha_{i_1 \cdots i_n} e_{i_1} \otimes \cdots \otimes e_{i_n} \right) \right\| \\
&\qquad \leqq \frac{n^n}{r_1^n} \sum |\lambda_{i_1} \cdots \lambda_{i_n} \alpha_{i_1 \cdots i_n}| \left\| \hat{\phi}_n \left(\frac{r_1}{n} e_{i_1} \otimes \cdots \otimes \frac{r_1}{n} e_{i_n} \right) \right\| \\
&\qquad \leqq \frac{n^n}{r_1^n} \frac{r_2}{n!} \sum |\lambda_{i_1} \cdots \lambda_{i_n} \alpha_{i_1 \cdots i_n}| \\
&\qquad \leqq \frac{n^n}{r_1^n} \frac{r_2}{n!} \|W\|_2^n \| \sum \alpha_{i_1 \cdots i_n} e_{i_1} \otimes \cdots \otimes e_{i_n} \|.
\end{aligned}$$

This implies that

$$\|\phi_n\| \leqq \frac{n^n}{r_1^n} \frac{r_2}{n!} \|W\|_2^n \leqq r_2 \frac{e^n}{r_1^n} \|W\|_2^n$$

for $n \geqq 2$ as required. □

Remark 3.4. (i) From the above proof it follows that $\hat{\alpha}_n$, where

$$\hat{\alpha}_1 := \|\phi'(0)\|$$

$$\hat{\alpha}_n := \frac{n^n r_2 \|W\|_2^n}{n! r_1^n} \quad \text{for } n \geqq 2$$

is a majorizing sequence for ϕ. In computations it turns out that it is easier to work with the majorizing sequence α_n given in the formulation of Lemma 3.3.

(ii) Note, moreover, we have that

$$\rho := \limsup (\alpha_n)^{1/n} = \frac{e\|W\|_2}{r_1}.$$

(iii) In what follows, we will assume that all of the input/output operators we consider are causal and are majorizable.

An interesting and useful property of fading memory operators is the following proposition.

PROPOSITION 3.5. *The notation and hypotheses are as in Lemma* 3.3. *Then each* ϕ_n (*regarded as a linear operator on* $G^{\otimes n}$) *is compact for* $n \geqq 2$.

Proof. Let the sequence in $G^{\otimes n}$

$$x^{(k)} := \sum_{i_1,\cdots,i_n=1}^{\infty} \alpha_{i_1\cdots i_n}^{(k)} e_{i_1} \otimes \cdots \otimes e_{i_n} \to 0$$

weakly. Define a projection in G for each natural number $N > 0$ by

$$P_N e_j = \begin{cases} 0 & j \leqq N \\ e_j & j \geqq N+1. \end{cases}$$

Then from the above proof of Lemma 3.3, for fixed n, we have that there exist constants C and $\hat{C}$ such that

$$\begin{aligned} \|\phi_n(x^{(k)})\| &\leqq C \sum_{i_1,i_2,\cdots,i_n \geqq 1} |\lambda_{i_1} \cdots \lambda_{i_n} \alpha_{i_1\cdots i_n}^{(k)}| \\ &\leqq C \sum_{i_1,i_2,\cdots,i_n \leqq N} |\lambda_{i_1} \cdots \lambda_{i_n} \alpha_{i_1\cdots i_n}^{(k)}| + \hat{C} \| WP_N \|_2 \| W \|_2^{n-1}. \end{aligned}$$

Thus,

$$\limsup \|\phi_n(x^{(k)})\| \leqq \hat{C} \| WP_N \|_2 \| W \|_2^{n-1}.$$

Hence as $N \to \infty$, we see that

$$\limsup \|\phi_n(x^{(k)})\| = 0,$$

which shows that ϕ_n is compact. □

4. Control theoretic preliminaries. We start here with the control problem definition. First, we will need to consider the precise kind of input/output operator we will be considering. See [3], [4] for closely related discussions. As mentioned above, we are assuming that all of the operators we consider are causal and are majorizable. For a discussion of causality in the nonlinear context, see [3]–[6]. Throughout this paper, $H^2(\mathbf{C}^k)$ will denote the standard Hardy space of $\mathbf{C}^k$-valued functions on the unit circle (k may be infinite, i.e., in this case $\mathbf{C}^k$ is replaced by h^2, the space of one-sided square-summable sequences). We now have the following definition.

DEFINITION 4.1. Let $S: H^2(\mathbf{C}^k) \to H^2(\mathbf{C}^k)$ denote the canonical unilateral right shift. Then we say an input/output operator ϕ is *locally stable* if it is causal and majorizable, $\phi(0) = 0$, and if there exists an $r > 0$ such that $\phi: B_r(H^2(\mathbf{C}^k)) \to H^2(\mathbf{C}^k)$ with $S\phi = \phi \circ S$ on $B_r(H^2(\mathbf{C}^k))$. We set

$$C_l := \{\text{space of locally stable operators}\}.$$

Since the theory we are considering is local, the notion of local stability is sufficient for all of the applications we have in mind. The interested reader can compare this notion with the more global notions of stability as, for example, discussed in [6].

The theory we are about to give holds for all plants which admit coprime locally stable factorizations. However, for simplicity we will assume that our plant is also locally stable. Accordingly, let P, W denote locally stable operators, with W invertible. Referring to Fig. 1, P represents the plant, and W the weight or filter. Now we say that the feedback compensator C *locally stabilizes* the closed loop if the operators

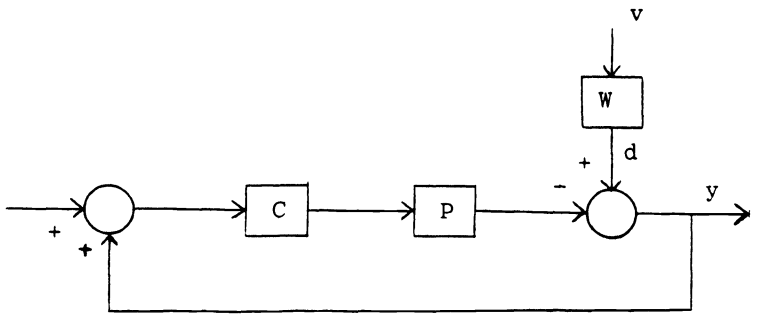

FIG. 1

$(I+P\circ C)^{-1}$ and $C\circ(I+P\circ C)^{-1}$ are well defined and locally stable. By a result of [2], C locally stabilizes the closed loop if and only if

$$C=\hat{q}\circ(I-P\circ\hat{q})^{-1} \tag{1}$$

for some $\hat{q}\in C_l$. Note then that the weighted sensitivity $(I+P\circ C)^{-1}\circ W$ can be written as $W-P\circ q$, where $q:=\hat{q}\circ W$. (Since W is invertible, the data q and $\hat{q}$ are equivalent.) In this context, we will call such a q a *compensating parameter.* From the compensating parameter q, we get a locally stabilizing compensator C via the formula (1).

The problem we would like to solve here is a version of the classical disturbance attenuation problem associated to the feedback loop in Fig. 1 (see [7], [21]). This, of course, corresponds to the "minimization" of the "sensitivity" $W-P\circ q$ taken over all locally stable q. In order to formulate a precise mathematical problem, we need to say in what sense we want to minimize $W-P\circ q$. This we will do in the next section where we will propose a notion of "sensitivity minimization" which seems quite natural to analytic input/output operators.

5. Sensitivity function. In this section we define a fundamental object, namely a nonlinear version of *sensitivity.* We will see that while the optimal H^∞ sensitivity is a real number in the linear case, the measure of performance which seems to be more natural in this nonlinear setting is a certain function defined in a real interval.

In order to define our notion of sensitivity, we will first have to partially order the space of analytic mappings defined in a ball about the origin. All of the input/output operators here will be locally stable. We also follow here our convention that for given $\phi\in C_l$, ϕ_n will denote the bounded linear map on the tensor space $(H^2(\mathbf{C}^k))^{\otimes n}$ associated with the n-linear part of ϕ, which we also denote by ϕ_n (and which we always assume without loss of generality is symmetric in its arguments). The context will always make the meaning of ϕ_n clear.

We can now state the following key definitions.

DEFINITION 5.1. (i) For $W, P, q\in C_l$ (W is the weight, P the plant, and q the compensating parameter), we define the *sensitivity functions* $S(q)$,

$$S(q)(\rho):=\sum_{n=1}\rho^n\|(W-P\circ q)_n\|$$

for all $\rho>0$ such that the sum converges. Note that for fixed P and W, for each $q\in C_l$, we get an associated sensitivity function.

(ii) We write $S(q)\lesssim S(\tilde{q})$, if there exists a $\rho_o>0$ such that $S(q)(\rho)\leqq S(\tilde{q})(\rho)$ for all $\rho\in[0,\rho_o]$. If $S(q)\lesssim S(\tilde{q})$ and $S(\tilde{q})\lesssim S(q)$, we write $S(q)\cong S(\tilde{q})$. This means that $S(q)(\rho)=S(\tilde{q})(\rho)$ for all $\rho>0$ sufficiently small, i.e., $S(q)$ and $S(\tilde{q})$ are equal as germs of functions.

(iii) If $S(q) \lesssim S(\tilde{q})$, but $S(\tilde{q}) \not\lesssim S(q)$, we will say that q *ameliorates* $\tilde{q}$. Note that this means $S(q)(\rho) < S(\tilde{q})(\rho)$ for all $\rho > 0$ sufficiently small.

Now with Definition 5.1, we can define a notion of "optimality" relative to the sensitivity function.

DEFINITION 5.2. (i) $q_o \in C_l$ is called *optimal* if $S(q_o) \lesssim S(q)$ for all $q \in C_l$.

(ii) We say $q \in C_l$ is *optimal with respect to its* nth *term* q_n, if for every n-linear $\hat{q}_n \in C_l$, we have

$$S(q_1 + \cdots + q_{n-1} + q_n + q_{n+1} \cdots) \lesssim S(q_1 + \cdots + q_{n-1} + \hat{q}_n + q_{n+1} + \cdots).$$

If $q \in C_l$ is optimal with respect to all of its terms, then we say that it is *partially optimal.*

Clearly, if q is optimal, then it is partially optimal; however, the converse may not hold. Note, moreover, that if ϕ is a Volterra series, then our definition of sensitivity measures in a precise sense the amplification of energy of each Volterra kernel on signals whose energy is bounded by a given ρ. For this reason, it appears that in this context, Definition 5.1 of the sensitivity function $S(q)$ seems physically natural. In the next section, we will discuss a procedure for constructing partially optimal compensating parameters, and then in § 7 we will show how this procedure leads to the construction of optimal compensating parameters for SISO systems. Of course, from formula (1) above, we can derive the corresponding partially optimal (respectively, optimal) compensator from the partially optimal (respectively, optimal) compensating parameter.

6. Iterative commutant lifting method. In this section, we discuss the main construction of this paper from which we will derive both partially optimal and optimal compensators relative to the sensitivity function given in Definition 5.1 above. As before, P will denote the plant and W the weighting operator, both of which we assume are locally stable. As in the linear case, we always suppose that P_1 is an isometry, i.e. P_1 is *inner.* In order to state our results, we will need to make a few preliminary remarks and set up some notation.

We begin by noting the following key relationship:

$$(W - P \circ q)_k = W_k - \sum_{1 \leqq j \leqq k} \sum_{i_1 + \cdots + i_j = k} P_j(q_{i_1} \otimes \cdots \otimes q_{i_j}).$$

Note that once again for ϕ majorizable, ϕ_n denotes the n-linear part of ϕ, as well as the associated linear operator on the appropriate tensor space.

We are now ready to formulate the *iterative commutant lifting procedure.* Let $\Pi : H^2(\mathbf{C}^k) \to H^2(\mathbf{C}^k) \ominus P_1 H^2(\mathbf{C}^k)$ denote orthogonal projection. Using the linear commutant lifting theorem (CLT) (see [19] for the details), we may choose q_1 such that

$$\| W_1 - P_1 q_1 \| = \| \Pi W_1 \|.$$

Now given this q_1, we choose (using the CLT) q_2 such that

$$\| W_2 - P_2(q_1 \otimes q_1) - P_1 q_2 \| = \| \Pi(W_2 - P_2(q_1 \otimes q_1)) \|.$$

Inductively, given $q_1, \cdots, q_{n-1}$, set

$$A_n := (W_n - \sum_{2 \leqq j \leqq n} \sum_{i_1 + \cdots + i_j = n} P_j(q_{i_1} \otimes \cdots \otimes q_{i_j})) \tag{2}$$

for $n \geqq 2$. Then from the CLT, we may choose q_n such that

$$\| A_n - P_1 q_n \| = \| \Pi A_n \|. \tag{2a}$$

We now come to the key point on the convergence of the iterative commutant lifting method.

PROPOSITION 6.1. *With the above notation, let* $q^{(1)} := q_1 + q_2 + \cdots$. *Then* $q^{(1)} \in C_l$.

Proof. It suffices to show that $\sum \|q_n\| \rho^n$ converges for all $0 \leqq \rho$ sufficiently small. Then from (2a)

$$\|A_n - P_1 q_n\| = \|\Pi A_n\| \leqq \|A_n\|$$

and so (using the fact that P_1 is an isometry)

$$\|q_n\| \leqq 2\|A_n\| \leqq 2\|W_n\| + 2 \sum_{2 \leqq j \leqq n} \sum_{i_1 + \cdots + i_j = n} \|P_j(q_{i_1} \otimes \cdots \otimes q_{i_j})\|. \tag{3}$$

Clearly from the majorizability hypothesis, we can find positive constants M_o, R_o, M, R such that

$$\|W_i\| \leqq \tfrac{1}{2} M_o R_o^i \tag{4}$$

$$\|P_j\| \leqq \tfrac{1}{2} M R^j \tag{5}$$

for $i \geqq 1$, and for $j \geqq 2$. Thus, $\|q_1\| \leqq M_o R_o$ and

$$\|q_n\| \leqq M_o R_o^n + \sum_{2 \leqq j \leqq n} M R^j \sum_{i_1 + \cdots + i_j = n} \|q_{i_1}\| \cdots \|q_{i_j}\| \tag{6}$$

for $n \geqq 2$. Let $f(z) = \sum_{n=0}^{\infty} f_n z^n$, and $g(z) = \sum_{n=0}^{\infty} g_n z^n$ be formal power series. Then we write $f \ll g$ if $|f_n| \leqq |g_n|$ for all $n \geqq 0$.

We introduce the notation

$$\tilde{q}(z) := \sum_{n=1}^{\infty} \|q_n\| z^n$$

$$a(z) := \sum_{n=1}^{\infty} M_o R_o^n z^n$$

$$b(z) := \sum_{n=2}^{\infty} M R^n z^n.$$

With this notation, (3) may be equivalently written as

$$\tilde{q}(z) \ll a(z) + b(\tilde{q}(z)). \tag{7}$$

Now (formally) define

$$\mu(z) \equiv a(z) + b(\mu(z)).$$

Then we claim the following:

(i) $\mu(z) \gg 0$;

(ii) $\tilde{q}(z) \ll \mu(z)$;

(iii) μ is analytic in some sufficiently small neighborhood of zero.

Clearly, the verification of this claim would complete the proof of the proposition.

In order to do this, let f be analytic in some ball of radius r_o centered at the origin. Then we set

$$\|f\|_{(r)} := \sup \{|f(z)| : |z| \leqq r\}$$

for $r < r_o$. Next, we define an operator on the set of analytic functions defined in some neighborhood of the origin by $F(f) := a + b(f)$ whenever $F(f)$ is well defined as an analytic function near zero. Then for given $\delta > 0$, and $r \leqq 1/2R_o < 1/R_o$ (this choice for r will be made clear below), we let

$$B := \{f \text{ analytic near } 0 : \|f - a\|_{(r)} \leqq \delta\}.$$

We want to choose δ, such that F is well defined in B, $F: B \to B$, and such that F is contractive in B.

Now it is easy to see that

$$\|f\|_{(r)} \leqq \delta + \|a\|_{(r)} \leqq \delta + \frac{M_o R_o r}{1 - R_o r} \leqq \delta + 2M_o R_o r.$$

Clearly, we can choose r, δ such that

$$0 < \delta + 2M_o R_o r \leqq \frac{1}{2R} < \frac{1}{R}. \tag{8a}$$

However,

$$\|F(f) - a\|_{(r)} \leqq \|b(f)\|_{(r)} \leqq \frac{MR^2(\delta + 2M_o R_o r)^2}{1 - (\delta + 2M_o R_o r)R} \leqq 2MR^2(\delta + 2M_o R_o r)^2. \tag{8b}$$

We require then that δ and r satisfy

$$2MR^2(\delta + 2M_o R_o r)^2 \leqq \delta. \tag{9}$$

With these choices we clearly have that $F: B \to B$. Now

$$\begin{aligned}
\|F(f) - F(g)\|_{(r)} \leqq \|b(f) - b(g)\|_{(r)} &\leqq \left\| \frac{M^2 f^2}{1 - Mf} - \frac{M^2 g^2}{1 - Mg} \right\|_{(r)} \\
&\leqq \left\| \frac{M^2(f^2 - g^2)}{1 - Mf} \right\|_{(r)} + M^2 \|g^2\|_{(r)} \left\| \frac{1}{1 - Mf} - \frac{1}{1 - Mg} \right\|_{(r)} \\
&\leqq 2M^2 \|f + g\|_{(r)} \|f - g\|_{(r)} + 4M^3 \|g\|_{(r)}^2 \|f - g\|_{(r)} \\
&\leqq (4M^2(\delta + 2M_o R_o r) + 4M^3(\delta + 2M_o R_o r)^2) \|f - g\|_{(r)}.
\end{aligned}$$

If we choose δ and r such that

$$\theta := (4M^2(\delta + 2M_o R_o r) + 4M^3(\delta + 2M_o R_o r)^2) < 1,$$

we see that

$$\|F(f) - F(g)\|_{(r)} \leqq \theta \|f - g\|_{(r)}.$$

Hence by the contraction mapping theorem, we get (iii). Moreover, (i) now follows immediately by definition of μ and the fact that $a(z) \gg 0$ and $b(z) \gg 0$. Finally, we can prove (ii) by induction. Indeed, let

$$\tilde{q}_k(z) := \sum_{n=1}^{k} \|q_n\| z^n$$

$$\mu_k(z) := \sum_{n=1}^{k} \mu_n z^n.$$

Clearly $\tilde{q}_1(z) \ll \mu_1(z)$, and suppose by induction that $\tilde{q}_n(z) \ll \mu_n(z)$ for $1 \leqq n \leqq N$. Then, note that there exists a polynomial p with positive coefficients depending on a and b such that $\tilde{q}_{N+1}(z) \ll p(\tilde{q}_1, \cdots, \tilde{q}_N)$ and $\mu_{N+1} = p(\mu_1, \cdots, \mu_N)$, from which (ii) follows immediately. This completes the proof of Proposition 6.1. □

Note that given any $q \in C_l$, we can apply the iterative commutant lifting procedure to $W - P \circ q$. Now set

$$S_\Pi(q)(\rho) := \sum_{n=1} \rho^n \|\Pi(W - P \circ q)_n\|.$$

Clearly, $S_\Pi(q) \subseteqq S(q)$ (as functions). We can now state the following result whose proof is immediate from the above discussion.

PROPOSITION 6.2. *Given $q \in C_l$, there exists $\tilde{q} \in C_l$, such that $S(\tilde{q}) \equiv S_\Pi(q)$. Moreover, $\tilde{q}$ may be constructed from the iterated commutant lifting procedure.*

Moreover, we easily have the following result.

PROPOSITION 6.3. *q is partially optimal if and only if $S(q) \cong S_\Pi(q)$ (i.e., $S(q)(\rho) = S_\Pi(q)(\rho)$ for all $\rho > 0$ sufficiently small; see § 5).*

Proof. Assume that q is partially optimal. Then, q must be optimal with respect to its first term q_1. However, we have seen that there exists $\hat{q}_1$ such that $\| W_1 - P_1 \hat{q}_1 \| = \| \Pi W_1 \|$. If $\| W_1 - P_1 q_1 \| > \| \Pi W_1 \|$, then since we are considering germs of functions, we would have $S(q) \not\cong S(\hat{q}_1 + q_2 + \cdots)$, contradicting the partial optimality of q.

By induction, assume that we have proven

$$\| (W - P \circ q)_j \| = \| \Pi (W - P \circ q)_j \|$$

for $1 \leqq j \leqq n$. Then again if

$$\| (W - P \circ q)_{n+1} \| > \| \Pi (W - P \circ q)_{n+1} \|,$$

by the above construction, using the commutant lifting theorem, we can find a $\hat{q}_{n+1}$ such that

$$\| \Pi (W - P \circ q)_{n+1} \| = \| (W - P \circ (q_1 + q_2 + \cdots + q_n + \hat{q}_{n+1} + \cdots))_{n+1} \|.$$

So once more, $S(q) \not\cong S(q_1 + \cdots + q_n + \hat{q}_{n+1} + q_{n+2} + \cdots)$, contradicting the partial optimality of q. Hence, we get that $S(q) \cong S_\Pi(q)$. The proof of the converse direction is similar. □

We can now summarize the above discussion with the following theorem.

THEOREM 6.4. *For given P and W as above, any $q \in C_l$ is either partially optimal or can be ameliorated by a partially optimal compensating parameter.*

Proof. The proof follows immediately from Propositions 6.1–6.3. □

It is important to emphasize that a partially optimal compensating parameter need not be optimal in the sense of Definition 5.1(i). Basically, what we have shown here is that using the iterated commutant lifting procedure, we can ameliorate any given design. The question of optimality will be considered in the next section.

7. Optimal compensators. In this section we will derive our main results about optimal compensators. Basically, we will show that in the single input/single output setting, the iterated commutant lifting procedure leads to an optimal design. We begin with the following theorem.

THEOREM 7.1. *There exist optimal compensators.*

Proof. We will only sketch the proof. Note that our proof is not constructive and makes use of the weak compactness property of weakly closed, bounded, convex sets of operators on Hilbert space.

First of all, set

$$O^{(1)} := \{q_1 : q_1 \text{ is optimal relative to } W_1 \text{ and } P_1\} = \{q_1 : \| \Pi W_1 \| = \| W_1 - P_1 q_1 \|\}.$$

It follows from the classical theory [1] that $O^{(1)}$ is a bounded, weakly closed, convex set of operators. Now set

$$\begin{aligned} O^{(2)}_{q_1} &:= \{q_2 : q_2 \text{ is optimal relative to } W_2 - P_2(q_1 \otimes q_1) \text{ and } P_1\} \\ &= \{q_2 : \| W_2 - P_2(q_1 \otimes q_1) - P_1 q_2 \| = \| \Pi (W_2 - P_2(q_1 \otimes q_1)) \|\}. \end{aligned}$$

Next let

$$\hat{W}_2(q) := W_2 - P_2(q \otimes q).$$

Further, we write

$$O^{(2)} := \bigcup_{q_1 \in O^{(1)}} O^{(2)}_{q_1}.$$

Then we can find a sequence $q_{2j} \in O^{(2)}_{q_{1j}}$ such that

$$\| \hat{W}_2(q_{1j}) - P_1 q_{2j} \| \to \inf \{ \| \hat{W}_2(q_1) - P_1 q_2 \| : q_1 \in O^{(1)}, q_2 \in O^{(2)} \} =: \sigma_2.$$

Without loss of generality, we can assume that $q_{1j} \to q_1$ weakly. Obviously $q_1 \in O^{(1)}$. Moreover, since $\{q_{2j}\}$ is a bounded sequence, we can also assume without loss of generality that $q_{2j} \to q_2$ weakly. Thus,

$$\| \hat{W}_2(q_1) - P_1 q_2 \| \leqq \liminf \| \hat{W}_2(q_{1j}) - P_1 q_{2j} \|,$$

and hence $\| \hat{W}_2(q_1) - P_1 q_2 \| = \sigma_2$.

Clearly the above procedure can be iterated step by step. Convergence follows by the same argument as that used in Proposition 6.1. □

For the construction of the optimal compensator in Theorem 7.3 below, we will need one more technical result. Accordingly, we will need to set up a bit more notation. First set $H^2 := H^2(\mathbf{C})$, and $H^\infty := H^\infty(\mathbf{C})$ (the space of bounded analytic complex-valued functions on the unit disc). Let $m \in H^\infty$ be a nonconstant inner function, let $\Pi_1 : H^2 \to H^2 \ominus mH^2 =: H(m)$ denote orthogonal projection, and set $T := \Pi_1 S | H(m)$, where S is the canonical unilateral shift on H^2. (T is the compressed shift.) For H a complex separable Hilbert space, let $S_\infty : H \to H$ denote a unilateral shift, i.e., an isometric operator with no unitary part. This means that $S_\infty^{*n} h \to 0$ for all $h \in H$ as $n \to \infty$. (See [15] and [19].) We can now state the following generalization of a nice result which appears in [18].

LEMMA 7.2. *Notation as above. Let $A : H \to H^2 \ominus mH^2$ be a bounded linear operator which attains its norm, i.e., such that there exists $h_0 \in H$ with $\|Ah_0\| = \|A\| \, \|h_0\| \neq 0$. Suppose moreover that*

$$AS_\infty = TA.$$

Then there exists a unique minimal intertwining dilation B of A, i.e., an operator $B : H \to H^2$ such that $BS_\infty = SB$, $\|A\| = \|B\|$, and $\Pi_1 B = A$.

Proof. First of all, without loss of generality, we can assume that $\|A\| = 1$. The existence of B follows from the commutant lifting theorem [19]. For the uniqueness, we use the results of [10]. Indeed, let

$$\mathbf{F} := \{ D_T Ah \oplus D_A h : h \in H \}^-$$

where for a contraction K, we set $D_K^2 := (I - K^*K)$, $D_K \geqq 0$. Then by [10], B is unique if $\mathbf{F} = \mathbf{D}_T \oplus \mathbf{D}_A$, where $\mathbf{D}_T = \overline{D_T H(m)}$, and $\mathbf{D}_A = \overline{D_A H}$. Now it is well known that $D_T f = \langle f, \mu \rangle \hat{\mu}$ where $\mu := \bar{z}(m(z) - m(0))$, and $\hat{\mu} := \mu / \|\mu\|$. Thus $D_T Ah = \langle Ah, \mu \rangle \hat{\mu}$, and so

$$\mathbf{F} = \{ \langle Ah, \mu \rangle \hat{\mu} \oplus D_A h : h \in H \}^-.$$

Since $h_0 \in H$ is such that $\|Ah_0\| = \|h_0\| \neq 0$, we have

$$D_T Ah_0 \oplus D_A h_0 = \langle Ah_0, \mu \rangle \hat{\mu} \oplus 0.$$

We consider the following two cases.

Case (i). Suppose $\langle Ah_0, \mu \rangle \neq 0$. Then $\mathbf{C}\hat{\mu} \oplus 0 \subset \mathbf{F}$, which implies that $\mathbf{F} \supset 0 \oplus \mathbf{D}_A$, from which we get that $\mathbf{F} = \mathbf{D}_T \oplus \mathbf{D}_A$.

Case (ii). Suppose $\langle Ah_0, \mu\rangle = 0$. We claim that there exists $j \geqq 1$ such that $\langle T^j Ah_0, \mu\rangle \neq 0$. Indeed, suppose not. Then $\langle T^j Ah_0, \mu\rangle = 0$ for all $j \geqq 1$; hence, $\|T^j Ah_0\| = \|Ah_0\| = \|h_0\|$. Let M be the Hilbert space generated by the elements $T^j Ah_0$ for $j \geqq 0$. Then M is T-invariant, and $T|M$ is an isometry. Since T is of class C_0 (see [19]), this is impossible. Thus, we can find a minimal j such that $\langle T^{j+1} Ah_0, \mu\rangle \neq 0$. However,

$$\|AS_\infty^j h_0\| = \|T^j Ah_0\| = \|Ah_0\| = \|h_0\|.$$

Hence replacing h_0 by $S_\infty^j h_0$, we are back to the first case, from which we can complete the proof. □

We now come to the main result of this section.

THEOREM 7.3. *Let W and P be* SISO *locally stable operators, with W the weight and P the plant. Suppose that ΠW_j is compact for $j \geqq 1$ and ΠP_k is compact for $k \geqq 2$. ($\Pi : H^2 \to H^2 \ominus P_1 H^2$ denotes orthogonal projection.) Let q_{opt} be a partially optimal compensating parameter as constructed by the iterated commutant lifting procedure. Then q_{opt} is optimal.*

Proof. First of all, since ΠW_1 attains its norm, from Lemma 7.2 we have that the optimal q_1 constructed relative to W_1 and P_1 is unique. (Actually, in this special case, since we are working in H^2, this follows from [18].) Now from our above hypotheses, each ΠA_k is compact for $k \geqq 2$; hence, each ΠA_k attains its norm. Therefore, by Lemma 7.2 each optimal q_k constructed by the iterated commutant lifting procedure is unique. Theorem 7.3 now follows immediately from Theorem 6.4. □

COROLLARY 7.4. *Let P be locally stable and* SISO, *with linear part P_1 rational. Then the partially optimal compensating parameter q_{opt} constructed by the iterated commutant lifting procedure is optimal.*

Proof. Indeed, since P_1 is SISO rational (recall that we also always assume that P_1 is inner), $H^2 \ominus P$, H^2 is finite-dimensional, and so we are done by Theorem 7.3. □

Remark 7.5. Corollary 7.4 gives a constructive procedure for finding the optimal compensator under the given hypotheses. Indeed, when P_1 is SISO rational, the iterative commutant lifting procedure can be reduced to *finite dimensional matrix calculations.* We will illustrate this important point via an example in § 9. In a subsequent paper, we will show that when the hypotheses of Theorem 7.3 are satisfied, the skew Toeplitz theory of [7] provides an algorithmic design procedure for distributed nonlinear systems as well.

8. Maximal vectors and optimal interpolants. In order to apply the iterative commutant lifting procedure to an actual example, we will need a generalization of a result due to Sarason [17] on the optimal interpolant. More precisely, for K a bounded linear operator on a Hilbert space, k_0 is a *maximal vector*, if $\|Kk_0\| = \|K\| \|k_0\| \neq 0$. Then for SISO systems, Sarason [17] derives a formula for the optimal interpolant in terms of a maximal vector of the associated Hankel operator (see [1] for a similar result).

In order to state our result, we will first need a few preliminary remarks. Let $H := H^2(\mathbf{C}^k)$. As above, we let $m \in H^\infty$ be nonconstant inner, and let $\Pi_1 : H^2 \to H^2 \ominus mH^2 =: H(m)$ denote orthogonal projection, with T the compression of the canonical shift on H^2 to $H(m)$. Moreover, S_∞ will denote the canonical shift on H, defined by multiplication by e^{it}. Now given $h \in H$, we can write h as a column vector (perhaps infinite)

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ \vdots \end{bmatrix}.$$

We then set

$$h^* := [\bar{h}_1\ \bar{h}_2 \cdots].$$

Moreover, given any bounded linear operator $B: H \to H^2$ such that $BS_\infty = SB$, we have that for $z \in D$ (the unit disc),

$$(Bh)(z) = \sum_{j \geq 1} b_j(z) h_j(z).$$

That is, we can express B as the row matrix

$$[b_1\ b_2 \cdots]$$

with $b_j \in H^\infty$ for $j \geq 1$. We will identify B with this row matrix. With this notation, we can now state the following resulting proposition.

PROPOSITION 8.1. *Notation as above. Let $A: H \to H^2 \ominus mH^2$ be a bounded linear operator such that $AS_\infty = TA$. Suppose, moreover, that A has a maximal vector h_0. Let $B: H \to H^2$ be the minimal intertwining dilation of A, i.e., $\Pi_1 B = A$, $BS_\infty = SB$, and $\|A\| = \|B\|$. Then if we let $\lambda := \|A\|^2$, we have that*

$$B = \frac{\lambda h_0^*}{\overline{Ah_0}}.$$

Proof. First of all, given $h_0 \in H$, we represent h_0 as a column vector with components h_j, $j \geq 1$ as above. Then, as we have seen, we have that $(Bh)(z) = \sum_{j \geq 1} b_j(z) h_j(z)$ (for $z \in D$), and

$$\|B\| = \sup\left\{\left(\sum_{j=1}^{\infty} |b_j(z)|^2\right)^{1/2} : |z| < 1\right\} = \operatorname{ess\,sup}\left\{\left(\sum_{j=1}^{\infty} |b_j(\zeta)|^2\right)^{1/2} : |\zeta| = 1\right\}.$$

However,

$$\|A\|^2 \|h_0\|^2 = \|Ah_0\|^2 \leq \|Bh_0\|^2 \leq \|B\|^2 \|h_0\|^2 = \|A\|^2 \|h_0\|^2.$$

Thus $\|Ah_0\|^2 = \|Bh_0\|^2$, and since $\Pi Bh_0 = Ah_0$, we have that $Ah_0 = Bh_0$. Next note that $\sum_{j \geq 1} |b_j(e^{it})|^2 \leq \lambda$ almost everywhere, and

$$\frac{1}{2\pi} \int_0^{2\pi} \left(\lambda \sum_{j=1}^{\infty} |h_j(e^{it})|^2 - \left|\sum_{j=1}^{\infty} b_j(e^{it}) h_j(e^{it})\right|^2\right) dt = 0.$$

(This follows from the fact that $\lambda \|h_0\|^2 = \|Bh_0\|^2$.) But using the Cauchy-Schwarz inequality, the expression under the integral sign is nonnegative. Thus,

$$\lambda \sum_{j \geq 1} |h_j(e^{it})|^2 = \left|\sum_{j \geq 1} b_j(e^{it}) h_j(e^{it})\right|^2 \leq \left(\sum_{j \geq 1} |b_j(e^{it})|^2\right)\left(\sum_{j \geq 1} |h_j(e^{it})|^2\right) \leq \lambda \sum_{j \geq 1} |h_j(e^{it})|^2$$

almost everywhere, which implies that

$$\sum_{j \geq 1} |b_j(e^{it})|^2 = \lambda$$

almost everywhere, and

$$h_j = \phi(e^{it})\overline{b_j(e^{it})}$$

almost everywhere for all $j \geq 1$, and for some function $\phi \in H^2$ satisfying

$$Ah_0 = Bh_0 = \lambda\phi.$$

Thus, for

$$B(e^{it}) = [b_1(e^{it}) b_2(e^{it}) \cdots]$$

we have

$$B(e^{it})\overline{Ah_0(e^{it})} = \lambda h_0(e^{it})^*$$

almost everywhere, as required. □

We will apply Proposition 8.1 in our computation of an optimal compensator in the next section.

9. Example. In this section, we will give an example of our nonlinear design procedure. Since we have been working in the disc, we will here take discrete-time systems, even though our techniques obviously go through in a similar manner for continuous-time systems as well. In what follows below, H_{D^2} will denote the space of $\mathbf{C}$-valued analytic functions on the bidisc D^2 with square integrable boundary values.

We let

$$W(z) = \frac{1-z}{2}$$

and $P = P_1 + P_2$, where $P_1 = z^2$ (in the discrete Fourier domain), and

$$P_2(F) = \frac{1}{2\pi i}\int_{|\zeta|=1} F(z\zeta^{-1}, \zeta)\frac{d\zeta}{\zeta}$$

for $F \in H_{D^2} \cong H^2 \otimes H^2$. More precisely, as we explained above, we can regard a bilinear map P_2 on $H^2 \times H^2$ as a linear map on $H^2 \otimes H^2$, and then it is easy to see that $H^2 \otimes H^2$ can be naturally identified with H_{D^2}. (The identification is given by $z \otimes 1 \to z_1$ and $1 \otimes z \to z_2$.) Note that in the discrete-time domain, P_2 is just a discrete Fourier transform of the "squaring" map, i.e., given the square integrable sequence $\{a_n\}$, we have that P_2 is the Fourier transform of the mapping $\{a_n\} \to \{a_n^2\}$.

We now apply our procedure to the weight W and the plant P. Accordingly, if we let $M_W : H^2 \to H^2$ denote multiplication by W, and let $\Pi : H^2 \to H^2 \ominus P_1 H^2 =: H_1$ be orthogonal projection, we set $A_0 := \Pi M_W | H_1$. Notice that $H_1 \cong \mathbf{C}^2$, and that via this isomorphism, we have the identification

$$A_0 = \begin{bmatrix} \frac{1}{2} & 0 \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

However,

$$A_0^* A_0 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

from which we get that $\|A_0\| = (\sqrt{5}+1)/2$, and that a maximal vector h_0 (i.e., a vector such that $\|A_0 h_0\| = \|A_0\| \, \|h_0\| \neq 0$) is given by

$$h_0 := \begin{bmatrix} 1 \\ -\beta \end{bmatrix}$$

where $\beta := (\sqrt{5}-1)/2$. Using then the Sarason formula [17] mentioned in the previous section, we can then compute that the optimal compensating parameter is

$$q_1 := \frac{\beta}{2(1-\beta z)}.$$

Of course, the above computation was based on standard linear H^∞-optimization theory. We now want to show how to get the optimal second-order compensating parameter. Accordingly, following the iterative commutant lifting procedure, we note that

$$P_2(q_1\otimes q_1)(F)=\frac{1}{2\pi i}\int_{|\zeta|=1} q_1(z\zeta^{-1})q_1(\zeta)F(z\zeta^{-1},\zeta)\frac{d\zeta}{\zeta}$$
$$=\frac{\beta^2}{8\pi i}\int_{|\zeta|=1}\frac{1}{1-\beta z\zeta^{-1}}\frac{1}{1-\beta\zeta}F(z\zeta^{-1},\zeta)\frac{d\zeta}{\zeta}$$

for $F\in H_{D^2}$. $P_2(q_1\otimes q_1)$ will be the "weight" for which we will apply the commutant lifting procedure relative to the "plant" P_1.

For $F\in H_{D^2}$, let

$$F(z_1,z_2)=\sum_{j,k=0}^{\infty}F_{jk}z_1^j z_2^k.$$

Then,

$$\frac{4}{\beta^2}P_2(q_1\otimes q_1)(F)=\sum_{j,k=0}^{\infty}\left(\frac{1}{2\pi i}\int_{|\zeta|=1}\frac{\zeta}{\zeta-\beta z}\frac{1}{1-\beta\zeta}z^j\zeta^{k-j}\frac{d\zeta}{\zeta}\right)F_{jk}$$
$$=\sum_{j,k=0}^{\infty}z^jF_{jk}\frac{(\beta z)^{k-j}}{1-\beta^2 z}+\sum_{j,k=0}^{\infty}z^jF_{jk}\frac{1}{2\pi i}\int_{0<|\zeta|=\varepsilon\ll 1}\frac{\zeta^{k-j}}{(\zeta-\beta z)(1-\beta\zeta)}d\zeta$$
$$=\sum_{j,k=0}^{\infty}F_{jk}\frac{\beta^{k-j}z^k}{1-\beta^2 z}+\sum_{j>k}\frac{z^jF_{jk}}{-\beta z}\frac{1}{2\pi i}$$
$$\times\int_{0<|\zeta|=\varepsilon\ll 1}\frac{(\sum_{l=0}^{\infty}\zeta^l/(\beta z)^l)(\sum_{l=0}^{\infty}\beta^l\zeta^l)}{\zeta^{j-k}}d\zeta$$
$$=\sum_{j,k=0}^{\infty}\frac{\beta^{k-j}z^k}{1-\beta^2 z}+\sum_{j>k}z^{j-1}F_{jk}\frac{-1}{\beta}\frac{1}{2\pi i}\int_{0<|\zeta|=\varepsilon\ll 1}\frac{\sum_{h,l}\beta^h(\zeta^{h+l}/(\beta z)^l)}{\zeta^{j-k}}d\zeta$$
$$=\sum_{j,k=0}^{\infty}F_{jk}\frac{\beta^{k-j}z^k}{1-\beta^2 z}+\sum_{j>k}\frac{-1}{\beta}F_{jk}\sum_{l+h=j-k-1}\beta^{h-l}z^{h+k}$$
$$=\sum_{j\leqq k}F_{jk}\frac{\beta^{k-j}z^k}{1-\beta^2 z}+\sum_{j>k}\beta^{k-j}F_{jk}z^k\frac{(z\beta^2)^{j-k}}{1-\beta^2 z}$$
$$=\sum_{j,k}F_{jk}\frac{\beta^{|k-j|}z^{\max\{j,k\}}}{1-\beta^2 z}.$$

Set $A:=-\Pi P_2(q_1\otimes q_1)$. Then from the above computations, we have that

$$-\frac{4}{\beta^2}AF=F_{00}+(\beta F_{10}+\beta F_{01})z+\beta^2F_{00}z+F_{11}z.$$

Moreover, if we let

$$A_1:=-\frac{4}{\beta^2}A|(\ker A)^\perp$$

we clearly have that

$$A_1\begin{bmatrix} F_{00} \\ F_{10} \\ F_{01} \\ F_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \beta^2 & \beta & \beta & 1 \end{bmatrix}\begin{bmatrix} F_{00} \\ F_{10} \\ F_{01} \\ F_{11} \end{bmatrix}$$

where we identify $(\ker A)^{\perp}$ with $\mathbf{C}^4$ in the natural way. Now

$$A_1A_1^* = \begin{bmatrix} 1 & \beta^2 \\ \beta^2 & (\beta^2+1)^2 \end{bmatrix},$$

and then it is easy to compute that $\|A_1\|^2 =: \lambda \cong 2.048924$, $\|A_1\| \cong 1.431406$, and that a maximal vector for A_1 is given by

$$h_1 := \begin{bmatrix} \lambda \\ (\lambda-1)/\beta \\ (\lambda-1)/\beta \\ (\lambda-1)/\beta^2 \end{bmatrix}.$$

Now we must write the Fourier representation of h_1 in order to apply Proposition 8.1, and so we must express H_{D^2} as some $H^2(\mathbf{C}^k)$. Accordingly, we apply the techniques of [19], to which we refer the reader for all the details about Fourier representations. More precisely, given $F = \sum_{j,k=0}^{\infty} F_{jk} z_1^j z_2^k$, we have that the Fourier representation of F, denoted by $F(\zeta)$, is given by

$$F(\zeta) := \sum_{n=0}^{\infty} \zeta^n \begin{bmatrix} F_{n,n} \\ F_{n+1,n} \\ F_{n,n+1} \\ F_{n+2,n} \\ F_{n,n+2} \\ \vdots \end{bmatrix} \tag{10}$$

for $\zeta \in \partial D$. Thus via the above identifications, the Fourier representation of h_1, denoted by $h_1(\zeta)$, is

$$h_1(\zeta) = \begin{bmatrix} \lambda \\ (\lambda-1)/\beta \\ (\lambda-1)/\beta \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} + \zeta \begin{bmatrix} (\lambda-1)/\beta^2 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix}.$$

Applying Proposition 8.1 (and using the same notation), we get that the minimal intertwining dilation of A_1, B_1, is given (in the Fourier space) by

$$B_1(\zeta) = \frac{\zeta\left[\lambda \dfrac{\lambda-1}{\beta} \dfrac{\lambda-1}{\beta} 0\,0\cdots\right] + \left[\dfrac{\lambda-1}{\beta^2} 0\,0\cdots\right]}{\zeta + (\lambda-1)/\beta^2}.$$

(Note that $(\lambda-1)/\beta^2$ is about $2.74>1$, and hence $1/(z+(\lambda-1)/\beta^2)$ is analytic and bounded in $\bar{D}$.) Using the Fourier representation (10) of F, we have that in the Fourier space

$$(B_1F)(\zeta)=B_1(\zeta)\sum_{n=0}^{\infty}\zeta^n\begin{bmatrix}F_{n,n}\\F_{n+1,n}\\F_{n,n+1}\\F_{n+2,n}\\F_{n,n+2}\\\vdots\end{bmatrix}$$

$$=\sum_{n=0}^{\infty}\zeta^n\frac{\zeta(\lambda F_{n,n}+(\lambda-1)\beta^{-1}F_{n+1,n}+(\lambda-1)\beta^{-1}F_{n,n+1})+(\lambda-1)\beta^{-2}F_{n,n}}{\zeta+(\lambda-1)\beta^{-2}}.$$

We are almost done! Indeed, still working with the Fourier representations, the optimal q_2 may be derived from the equality (for $z\in D$)

$$-(4/\beta^2)P_2(q_1\otimes q_1)F-z^2q_2F=-B_1F.$$

Thus, we see that

$$\begin{aligned}(q_2F)(z)&\\&=\frac{1}{z^2}\frac{(\sum_{n=0}^{\infty}z^n\{z(\lambda F_{n,n}+(\lambda-1)\beta^{-1}F_{n+1,n}+(\lambda-1)/\beta^{-1}F_{n,n+1})+(\lambda-1)\beta^{-2}F_{n,n}\})}{z+(\lambda-1)\beta^{-2}}\\&\quad-\frac{1}{z^2}\sum_{j,k}F_{jk}\frac{\beta^{|k-j|}z^{\max\{j,k\}}}{1-\beta^2z}.\end{aligned}\tag{11}$$

Despite its seemingly complicated form, we will now see that q_2 has an integral expression in the Fourier domain, which translates into a rather simple two-linear function in the time-domain. Explicitly, we may write (11) equivalently as

$$\begin{aligned}q_2=S_1-S_2+\frac{\lambda F_{1,1}+(\lambda-1)\beta^{-1}(F_{2,1}+F_{1,2})}{z+(\lambda-1)\beta^{-2}}-\frac{\lambda F_{1,1}}{(1-\beta^2z)(z+(\lambda-1)\beta^{-2})}\\-\frac{\beta^2\lambda F_{0,0}+\beta\lambda(F_{1,0}+F_{0,1})}{(1-\beta^2z)(z+(\lambda-1)\beta^{-2})},\end{aligned}\tag{12}$$

where

$$S_1:=\frac{(\sum_{n=2}^{\infty}z^{n-2}\{z(\lambda F_{n,n}+(\lambda-1)/\beta F_{n+1,n}+(\lambda-1)/\beta F_{n,n+1})+(\lambda-1)/\beta^2F_{n,n}\})}{z+(\lambda-1)/\beta^2}$$

and

$$S_2:=\sum_{n=2}^{\infty}\frac{z^{n-2}}{1-\beta^2z}\left(\sum_{0\leqq j\leqq n}(F_{jn}+F_{nj})\beta^{n-j}\right).$$

Clearly in order to find a computable expression for q_2, we must first find such an expression for the map $M_{m,n}:H_{D^2}\to\mathbf{C}$, defined by $M_{m,n}(F):=F_{m,n}$ where $m,n=0,1,\cdots$ are fixed. Let $a=\{a_j\}$ and $b=\{b_j\}$ $(j\geqq 0)$ be sequences in the "discrete time-domain" h^2. By slight abuse of notation, we also let $a=a(\zeta)=\sum_{j=0}^{\infty}a_j\zeta^j$, and $b=b(\zeta)=\sum_{j=0}^{\infty}b_j\zeta^j$ denote their discrete Fourier transforms. Then it is easy to see that

$$M_{m,n}(a\otimes b)=M_{m,n}(a,b)=\left(\frac{1}{2\pi i}\right)^2\int_{|\zeta_1|=1}\int_{|\zeta_2|=1}\zeta_1^{-m}\zeta_2^{-n}a(\zeta_1)b(\zeta_2)\frac{d\zeta_1}{\zeta_1}\frac{d\zeta_2}{\zeta_2}=a_mb_n.$$

In this way, we get that

$$S_1 = \frac{1}{z+(\lambda-1)/\beta^2}\, T_1(F)$$

where

$$T_1(F) := \sum_{n=2}^{\infty} z^{n-2}\{z(\lambda F_{n,n}+(\lambda-1)/\beta F_{n+1,n}+(\lambda-1)/\beta F_{n,n+1})+(\lambda-1)/\beta^2 F_{n,n}\}.$$

Hence, we see that

$$T_1(a\otimes a) = \lambda \sum_{n=2}^{\infty} z^{n-1}a_n^2 + 2(\lambda-1)\beta^{-1}\sum_{n=2}^{\infty} z^{n-1}a_n a_{n+1} + (\lambda-1)\beta^{-2}\sum_{n=2}^{\infty} z^{n-2}a_n^2$$

which, of course, is the transform of a very simple quadratic map in the time domain. In the exact same way, we can write down explicit expressions for all the terms of q_2 appearing in formula (12).

Note that our above computations essentially amount to finite-dimensional matrix manipulations. We have then that q_1+q_2 is the optimal compensating parameter up to order two. A similar computation allows us to find the optimal compensating parameter up to any order, and by Proposition 6.1, our procedure is guaranteed to converge.

10. Conclusions. In this paper we have introduced a novel notion of "sensitivity minimization," and have given a method for constructing optimal compensators for SISO systems, and partially optimal compensators for MIMO systems. This generalizes the standard H^∞ linear theory in a rather natural way. However, in contrast to the linear case, the measure of performance is now given by (the germ of) a certain sensitivity function instead of a real number. The key idea is the utilization of an iterative commutant lifting procedure which can also be employed to ameliorate any given design in the sense of § 5.

The techniques we have used here are local and very much inspired by the previous work in [3]–[5]. The interested reader can contrast this approach with the nonlinear Ball–Helton method as given in [6]. An intriguing problem would be to compare nonlinear designs derived from these two approaches (which, of course, coincide in the linear case). This we would like to consider in some future work as well as attempt to derive a more global theory. There are, of course, a number of open questions still remaining even in our local setting. A key problem is to design optimal controllers for nonlinear MIMO plants. Indeed, even though we can ameliorate any design, because of nonuniqueness in the choice of the various minimal intertwining dilations in the iterative commutant lifting procedure, for MIMO systems we cannot guarantee optimality but only partial optimality. In a subsequent paper, we plan to show how the skew Toeplitz techniques of [7] provide a design methodology for distributed nonlinear systems as well.

At the Systems Research Center of Honeywell in Minneapolis, an interesting partial dynamic inversion technique due to Elgersma and Morton [9] has recently been employed to obtain some nonlinear designs related to a sixth degree of freedom aircraft model. A project on which we are now embarked is the utilization of the iterative commutant lifting procedure in order to ameliorate this kind of design. Finally, in the SISO case (in which there is a rather complete theory), our procedure is algorithmic, and we are presently working on software for its digital implementation with our colleagues at Honeywell along the lines of the work already done in the linear framework based on [11] and [12].

REFERENCES

[1] V. M. ADAMJAN, D. Z. AROV, AND M. G. KREIN, *Infinite Hankel matrices and generalized problems of Caratheodory-Fejer and F. Riesz*, Functional Anal. Appl., 2 (1968), pp. 1-18.

[2] V. ANANTHARAM AND C. DESOER, *On the stabilization of nonlinear systems*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 569-573.

[3] J. BALL, C. FOIAS, J. W. HELTON, AND A. TANNENBAUM, *On a local nonlinear commutant lifting theorem*, Indiana J. Mathematics, 36 (1987), pp. 693-709.

[4] ———, *Nonlinear interpolation theory in H^∞*, in Modelling, Robustness, and Sensitivity in Control Systems, R. Curtin, ed., NATO-ASI Series, Springer-Verlag, New York, 1987.

[5] ———, *A Poincare-Dulac approach to a nonlinear Beurling-Lax-Halmos theorem*, J. Math. Anal. Appl., to appear.

[6] J. BALL AND J. W. HELTON, *Sensitivity bandwidth optimization for nonlinear feedback systems*, Technical Report, Department of Mathematics, University of California at San Diego, 1988.

[7] H. BERCOVICI, C. FOIAS, AND A. TANNENBAUM, *On skew Toeplitz operators*, I, Operator Theory: Adv. Appl., 29 (1988), pp. 21-44.

[8] S. BOYD AND L. CHUA, *Fading memory and the problem of approximating nonlinear operators with Volterra series*, IEEE Trans. Circuits and Systems, CAS-32 (1985), pp. 1150-1161.

[9] M. ELGERSMA AND B. MORTON, *Nonlinear flying quality parameters based on dynamic inversion*, Technical Report AFWAL-TR-87-3079, 1987.

[10] C. FOIAS AND A. FRAZHO, *On the Schur representation in the commutant lifting theorem*, I, Operator Theory: Adv. Appl., 18 (1986), pp. 207-217.

[11] C. FOIAS AND A. TANNENBAUM, *On the four block problem*, II: *the singular system*, Integral Equations Operator Theory, 11 (1988), pp. 726-767.

[12] 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. 1081-1089.

[13] B. FRANCIS, *A Course in H^∞ Control Theory*, McGraw-Hill, New York, 1987.

[14] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications XXIII, Providence, 1957.

[15] M. ROSENBLUM AND J. ROVNYAK, *Hardy Classes and Operator Theory*, Oxford University Press, New York, 1985.

[16] W. J. RUGH, *Nonlinear System Theory: the Volterra/Wiener Approach*, Johns Hopkins Univ. Press, Baltimore, 1981.

[17] D. SARASON, *Generalized interpolation in H^∞*, Trans. Amer. Math. Soc., 127 (1967), pp. 179-203.

[18] ———, *Function Theory on the Unit Circle*, Lecture Notes, Virginia Polytechnic Institute, Blacksburg, VA, 1978.

[19] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.

[20] N. WIENER, *Nonlinear Problems in Random Theory*, Technology Press of M.I.T., Cambridge, MA, 1958.

[21] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 301-320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 861-875, July 1989

011

NONLINEAR OPTIMAL CONTROL WITH INFINITE HORIZON FOR DISTRIBUTED PARAMETER SYSTEMS AND STATIONARY HAMILTON–JACOBI EQUATIONS*

P. CANNARSA† AND G. DA PRATO‡

Abstract. Optimal control problems, with no discount, are studied for systems governed by nonlinear "parabolic" state equations, using a dynamic programming approach.

If the dynamics are stabilizable with respect to cost, then the fact that the value function is a generalized viscosity solution of the associated Hamilton-Jacobi equation is proved. This yields the feedback formula. Moreover, uniqueness is obtained under suitable stability assumptions.

Key words. optimal control, Hamilton-Jacobi equations, viscosity solutions, evolution equations, unbounded operators

AMS(MOS) subject classifications. 49C20, 34G20

1. Introduction and setting of the problem. Let us consider two separable reflexive Banach spaces, X (the *state space*) and U (the *control space*). We denote by $|\ |$ the norm of X, which we assume to be continuously differentiable in $X\backslash\{0\}$, by X^* the dual space of X and by $\langle\ ,\ \rangle$ the pairing between X and X^*. We denote by $\partial|x|$ the subgradient of $|x|$, which is obviously single-valued on $X\backslash\{0\}$. The same symbols will also be used in the Banach space U. Moreover, we will use the following notation:

(i) For any Banach space K and any nonnegative integer k we denote by $C^k(X; K)$ the set of all the mappings $f: X \to K$ that are continuous and bounded on all bounded sets of X, together with their derivatives of order less than or equal to k.

(ii) We denote by $C^{k,1}(X; K)$ (respectively, $C^{k,1}(X; K)_{\text{loc}}$), the set of all the mappings f in $C^k(X; K)$ whose derivative of order k is Lipschitz continuous in X (in every bounded set of X).

We are interested in the following optimal control problem.

Minimize

$$J_\infty(u, x) = \int_0^\infty \{g(y(s)) + h(u(s))\}\, ds \tag{1.1}$$

over all $u \in L^1(0, \infty; U)_{\text{loc}}$, subject to state equation

$$y' = Ay + F(y) + Bu, \qquad y(0) = x. \tag{1.2}$$

Following the dynamic programming approach, we will study the Hamilton-Jacobi equation

$$H(B^*DV(x)) - \langle Ax + F(x), DV(x)\rangle - g(x) = 0 \tag{1.3}$$

where H denotes the Legendre transform of h, that is,

$$H(v) = \sup_{v \in U} \{-\langle u, v\rangle - h(u)\}. \tag{1.4}$$

The connections between (1.3) and problem (1.1)-(1.2) are well known.

* Received by the editors July 15, 1988; accepted for publication (in revised form) November 7, 1988. This research was supported by Consiglio Nazionale delle Ricerche.

† Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti, 2, 56127 Pisa, Italy.

‡ Scuola Normale Superiore, 56126 Pisa, Italy.

We assume the following hypotheses.

(SL) (i) $A: D(A) \subset X \to X$ generates an analytic semigroup e^{tA} in X and there exists $\omega \in \mathbf{R}$ such that $\|e^{tA}\| \leqq e^{\omega t}$.

(ii) The embedding $D(A) \to X$ is compact.

(iii) $B \in \mathscr{L}(U; X)$.

(iv) $F \in C^{1,1}(X, X)_{\text{loc}}$ and there exists $a \in \mathbf{R}$ such that $\langle F(x), x^* \rangle \leqq a|x|$, for all $x^* \in \partial|x|$, for all $x \in X$.

(v) $g \in C^{1,1}(X, \mathbf{R})_{\text{loc}}$ and $g(x) \geqq 0$ for all $x \in X$.

(vi) $h \in C^{1,1}(U, \mathbf{R})_{\text{loc}}$ is strictly convex and there exists $p > 1$ such that $h(u) \geqq \gamma|u|^p$ for all $u \in U$ and some $\gamma > 0$.

We remark that, if (SL) are fulfilled, then, by classical arguments (see, for instance, [21]), problem (1.2) has a unique global mild solution $y \in C([0, \infty[; X)$.

In the analysis of (1.3), we meet with two immediate difficulties: the nonsmoothness of solutions and the unboundedness of A. In fact, first-order partial differential equations have, in general, no global classical solutions even in finite dimensions. Therefore, a suitable notion of weak solution is required. Moreover, such a generalized solution will have to take care of the fact that Ax is defined only on a dense subspace of X.

The first problem can be successfully treated by the notion of viscosity solution, introduced by Crandall and Lions [12]-[15]. In [16] they have also extended their definition of solutions to problems involving unbounded operators. Further results in these directions have been obtained in [4] and [10] by an approximation procedure.

Stationary Hamilton-Jacobi equations have been extensively studied (see [20] for general references and results; see also [13]-[16]) mainly in the case when the equation contains an additional term of the form λV with $\lambda > 0$. This corresponds to the introduction of a discount factor $e^{-\lambda t}$ in the cost.

However, in many applications we are required not to have such a discount, as in linear quadratic optimal control problems. A large amount of work has been devoted to the analysis of this case (see, for instance, the review paper [22]). For linear quadratic optimal control problems, the Hamilton-Jacobi equation is replaced by the algebraic Riccati equation, as it is well known. In general, uniqueness is false for this equation. Therefore, we do not expect to have uniqueness for (1.3).

Optimal control problems with a linear state equation and a convex cost functional are also studied in [3] and [6]. Some generalizations to the nonconvex case, by using variational methods, are contained in [4], [6], and [7].

The main idea of our approach is to obtain a viscosity solution V of (1.3) as

$$V(x) = \lim_{t \to \infty} \phi(t, x) \tag{1.5}$$

where ϕ solves the forward equation (in the generalized sense of [10]):

$$\phi_t(t, x) + H(B^* \nabla \phi(t, x)) - \langle Ax + F(x), \nabla \phi(t, x) \rangle - g(x) = 0, \qquad \phi(0, x) = 0. \tag{1.6}$$

For the value function of the control problem (1.1)-(1.2) to be finite, we introduce the notion of stabilizability that generalizes a well-known concept in linear quadratic control (see, e.g., [22]).

DEFINITION 1.1. We say that $(A + F, B, h)$ is *stabilizable with respect to the observation* g (or, for brevity, that problem (1.1)-(1.2) is stable) if for any $x \in X$ there exists $u_x \in L^1(0, \infty; U)_{\text{loc}}$ such that $J_\infty(u_x, x) < \infty$. Such a control u_x will be called an *admissible control* at x.

Finally, we define the *value function* of problem (1.1) (1.2) as

$$V_\infty(x) = \inf\{J_\infty(u, x);\ u \in L^1(0, \infty; U)_{\text{loc}}\}. \tag{1.7}$$

We say that $u^* \in L^1(0, \infty; U)_{\text{loc}}$ is an *optimal control* if $J(u^*) = V_\infty(x)$; in this case, we call the corresponding solution y^* of (1.2) an *optimal state* and (u^*, y^*) an *optimal pair* at x.

In this paper we show that, if $(A+F, B, h)$ is g-stabilizable, then V_∞ is a generalized viscosity solution of (1.3). Moreover, we obtain the existence of optimal pairs as well as the feedback formula (see Theorem 4.4).

In § 3 we study the "stability" of the closed-loop system. When this system is stable and B is invertible, we prove the uniqueness of the nonnegative generalized viscosity solution of (1.3) vanishing at zero (Theorem 5.4).

An application to a nonlinear control problem for a distributed parameter system is illustrated in § 6.

We now explain our definition of generalized solutions. We define solutions of (1.3) as stationary solutions of the following evolution equation:

$$-W_t(t, x) + H(B^*\nabla W(t, x)) - \langle Ax + F(x), \nabla W(t, x)\rangle = g(x). \tag{1.8}$$

More precisely, we have the following definition.

DEFINITION 1.2. Assume (SL). We say that $V \in C^{0,1}(X; \mathbf{R})_{\text{loc}}$ is a *generalized viscosity solution* of (1.3) if $W(t, x) := V(x)$ is the generalized viscosity solution of (1.8) in $[0, T] \times X$ with terminal data $W(T, x) = V(x)$, for all $T > 0$.

We recall below the definition of generalized viscosity solutions of the Cauchy problem (see [10])

$$\begin{aligned} &-W_t(t, x) + H(B^*\nabla W(t, x)) - \langle Ax + F(x), \nabla W(t, x)\rangle = g(x) \\ &W(T, x) = \phi_0(x), \qquad x \in X, \quad t \in [0, T] \end{aligned} \tag{1.9}$$

where

$$\phi_0 \in C^{0,1}(X; \mathbf{R})_{\text{loc}}. \tag{1.10}$$

DEFINITION 1.3. Assume (SL) and (1.10). We say that $W \in C([0, T] \times X; \mathbf{R})$ is a *generalized viscosity solution* of (1.9) if we have

$$\lim_{n\to\infty} W_n(t, x) = W(t, x), \quad \forall x \in D(A), \quad \forall t \in [t, T] \tag{1.11}$$

where W_n is the viscosity solution (in the sense of Crandall and Lions [13]) of the problem

$$\begin{aligned} &-W_{nt}(t, x) + H(B^*\nabla W_n(t, x)) - \langle A_n x + F(x), \nabla W_n(t, x)\rangle - g(x) = 0, \\ &W_n(T, x) = \phi_0(x) \end{aligned} \tag{1.12}$$

where

$$A_n = nA(n - A)^{-1}. \tag{1.13}$$

We note that problem (1.12) has a unique viscosity solution (see [13] and also [10]).

A property of generalized viscosity solutions that turns out to be essential to our approach is semiconcavity (see [9]).

In applications it is also useful to consider the following more general assumptions:

(SL') (i) Hypotheses (SL) (i), (ii), (iii), (v) and (vi) hold.

(ii) there exists a Banach space Z (with pairing denoted $\langle\ ,\ \rangle_Z$), continuously embedded in X, such that the part of A in Z, A_Z, generates an analytic semigroup in Z with domain $D(A_Z)$ (not necessarily dense in Z)

$$D(A_Z)=\{x\in D(A)\cap Z;\ Ax\in Z\}.$$

Moreover, $\|e^{tA_Z}\|\leqq e^{\mu t}$ for all $t\geqq 0$ and some $\mu\in\mathbf{R}$.

(iii) There exists $\alpha\in]0,1-1/p[$, $a\in\mathbf{R}$, and two continuous functions β, $\rho:[0,\infty[\to[0,\infty[$, such that $D_A(\alpha,p)$ is embedded in Z and

$$F\in C^{1,1}(D_A(\alpha,p);X)_{\mathrm{loc}}, \tag{1.14}$$

$$\langle F(z),z^*\rangle_Z\leqq a|z|_Z\quad\forall z\in Z,\quad\forall z^*\in\partial|z|_Z, \tag{1.15}$$

$$|F(x)|\leqq\beta(|x|_Z)+\rho(|x|_Z)|x|_{\alpha,p}\quad\forall x\in D_A(\alpha,p). \tag{1.16}$$

We recall that $D_A(\alpha,p)$ is the real interpolation space between $D(A)$ and X, introduced by Lions and Peetre [19], with norm

$$|x|_{\alpha,p}=\left[\int_0^\infty \tau^{p-p\alpha-1}|Ae^{\tau A}x|^p\,d\tau\right]^{1/p}.$$

Definition 1.2 remains unchanged under assumptions (SL'), except for the fact that we assume $V\in C^{0,1}(D_A(\alpha,p);\mathbf{R})_{\mathrm{loc}}$. Moreover, in Definition 1.3 we assume $W\in C([0,T]\times D_A(\alpha,p);\mathbf{R})$ and replace (1.12) by

$$-W_{nt}(t,x)+H(B^*\nabla W_n(t,x))-\langle A_nx+F(n(n-A)^{-1}x),\nabla W_n(t,x)\rangle-g(x)=0,$$

$$W_n(T,x)=\phi_0(x).$$

2. Preliminaries. In this section we recall the basic results on the time-dependent Hamilton–Jacobi equation (1.9).

PROPOSITION 2.1. *Assume* (1.10) *and either* (SL) *or* (SL'). *Then, there exists a unique generalized viscosity solution* W *of problem* (1.9) *given by*

$$W(t,x)=\inf\left\{\int_t^T[g(y(s))+h(u(s))]\,ds+\phi_0(y(T));\quad u\in L^1(t,T;U)_{\mathrm{loc}}\right\} \tag{2.1}$$

where y *is the solution of*

$$y'(s)=Ay(s)+F(y(s))+Bu(s),\qquad t\leqq s\leqq T,\quad y(t)=x. \tag{2.2}$$

Moreover, W *satisfies* (1.9) *in the sense that for every* $(t,x)\in[0,T]\times D(A)$ *we have*

$$\text{(i)}\quad\forall(p_t,p_x)\in D^+W(t,x),\qquad -p_t+H(B^*p_x)-\langle Ax+F(x),p_x\rangle\leqq g(x), \tag{2.3}$$

$$\text{(ii)}\quad\forall(p_t,p_x)\in D^-W(t,x),\qquad -p_t+H(B^*p_x)-\langle Ax+F(x),p_x\rangle\geqq g(x).$$

We recall the definition of the semidifferentials D^+ and D^-:

$$D^+W(t,x)=\left\{(p_t,p_x)\in\mathbf{R}\times X^*;\ \limsup_{(s,y)\to(t,x)}\frac{W(s,y)-W(t,x)-(s-t)p_t-\langle y-x,p_x\rangle}{|s-t|+|y-x|}\leqq 0\right\},$$

$$D^-W(t,x)=\left\{(p_t,p_x)\in\mathbf{R}\times X^*;\ \liminf_{(s,y)\to(t,x)}\frac{W(s,y)-W(t,x)-(s-t)p_t-\langle y-x,p_x\rangle}{|s-t|+|y-x|}\geqq 0\right\}. \tag{2.4}$$

Remark 2.2. The results of Proposition 2.1 are proved in Theorems 3.3 and 3.7 of [10] in a slightly different form that is equivalent to the one above in view of the coercivity assumption on h.

We now recall the Maximum Principle [8], the feedback formula [4], [9], and some regularity properties of optimal pairs [9].

PROPOSITION 2.3. *Assume* (SL) (*respectively*, (SL')) *and* (1.10). *Let W be given by* (2.1) *and* $(t, x) \in [0, T] \times X$ (*respectively*, $(t, x) \in [0, T] \times D_A(\alpha, p)$). *Let* (u^*, y^*) *be an optimal pair for W at* (t, x). *Then, there exists* $p^* \in C([t, T]; X^*)$ *such that*

$$p^{*\prime}(s) + A^* p^*(s) + (DF(y^*(s))^* p^*(s) + Dg(y^*(s)) = 0, \quad p^*(T) = D\phi(y^*(T)), \tag{2.5}$$

$$u^*(s) = -DH(B^* p^*(s)), \qquad t \leqq s \leqq T. \tag{2.6}$$

We call p^* *a* <u>*dual arc*</u>. *Moreover,*

$$u^*(s) \in -DH(B^* \nabla^+ W(s, y^*(s))), \qquad t \leqq s \leqq T \tag{2.7}$$

where

$$\nabla^+ W(s, x) = \left\{ q \in X^*; \limsup_{y \to x} \frac{W(s, y) - W(s, x) - \langle y - x, q \rangle}{|y - x|} \leqq 0 \right\}. \tag{2.8}$$

Furthermore, there exists $\delta \in]0, 1[$ *such that*

$$y^* \in C^{1,\delta}(]t, T[; X), \tag{2.9}$$

$$p^* \in C^{1,\delta}(]t, T[; X), \qquad u^* \in C^{0,\delta}(]t, T[; X). \tag{2.10}$$

Above we have denoted by $C^{1,\delta}(I; X)$, for any real interval I, the space of functions that are Hölder continuous with exponent δ, together with their first derivative, on each subinterval $[a, b]$ contained in I.

Finally, the following results are proved in [4] and [9].

PROPOSITION 2.4. *Assume* (SL) (*respectively*, (SL')) *and* (1.10) *and let W be given by* (2.1). *Then we have the following*:

(2.11) (i) $W(t, .)$ *is locally Lipschitz in X for all* $t \in [0, T]$;
(ii) $W(., x)$ *is Lipschitz continuous in* $[0, T]$ *for all* $x \in D(A)$.

Furthermore, if $B^{-1} \in \mathcal{L}(H; U)$, *then* $W(t, .)$ *is semiconcave in X for all* $t \in [0, T[$; *that is, for all* $r > 1/T$ *there exists* $C_r > 0$ *such that*

$$\lambda W(t, x + (1 - \lambda)x') + (1 - \lambda) W(t, x - \lambda x') - W(t, x) \leqq C_r \lambda (1 - \lambda)|x'|^2$$

for all $t \in [0, T - 1/r]$, $|x|, |x'| \leqq r$, $\lambda \in [0, 1]$.

Along with the backward Cauchy problem (1.9), we will consider the forward problem:

$$\begin{gathered} \phi_t(t, x) + H(B^* \nabla \phi(t, x)) - \langle Ax + F(x), \nabla \phi(t, x) \rangle - g(x) = 0; \\ \phi(0, x) = \phi_0(x). \end{gathered} \tag{2.12}$$

We say that $\phi \in C([0, T] \times X; \mathbf{R})$ (respectively, $\phi \in ([0, T] \times D_A(\alpha, p); \mathbf{R})$) is the *generalized viscosity solution* of (2.12) if $W(t, x) = \phi(T - t, x)$ is the generalized viscosity solution of (1.9).

We prove now the analogue of representation formula (2.1).

PROPOSITION 2.5. *Assume* (SL) (*respectively*, (SL')) *and* (1.10). *Let* ϕ *be the generalized viscosity solution of* (2.12). *Then we have*

$$\phi(t, x) = \inf \left\{ \int_0^t [g(y(s)) + h(u(s))]\, ds + \phi_0(y(t)); u \in L^1(0, \infty; U)_{\text{loc}} \right\} \tag{2.13}$$

where y is the solution of (1.2).

Proof. By definition we have

$$\phi(t,x)=\inf\Big\{\int_{T-t}^{T}\{g(y(s))+h(u(s))\}\,ds+\phi_0(y(T));$$

$$u\in L^1(T-t,T;U),\ y'(s)=Ay(s)+F(y(s))+Bu(s),\ y(T-t)=x\Big\}. \tag{2.14}$$

Set $\sigma=s-T+t$ to obtain

$$\phi(t,x)=\inf\Big\{\int_0^t g(y(\sigma+T-t))+h(u(\sigma+T-t))\,d\sigma+\phi_0(y(T));$$

$$u\in L^1(T-t,T:U),\ y'(s)=Ay(s)+F(y(s))+Bu(s),\ y(T-t)=x\Big\}.$$

Now, let $\underline{y}(s)=y(s+T-t)$, $\underline{u}(s)=u(s+T-t)$; then

$$\phi(t,x)=\inf\Big\{\int_0^t\{g(\underline{y}(s))+h(\underline{u}(s))\}\,ds+\phi_0(\underline{y}(t));$$

$$\underline{u}\in L^1(0,t;U),\ \underline{y}'(s)=A\underline{y}(s)+F(\underline{y}(s))+B\underline{u}(s),\ \underline{y}(0)=x\Big\}$$

and the assertion is proved. □

3. Sufficient conditions for stabilizability. To our knowledge there are no general conditions that yield global stabilizability in the sense of Definition 1.1 (for local results see [2] and [18]). In the following we give some sufficient conditions that may be applied to various situations. For instance, the problem we analyze in § 6 fits into the framework of Proposition 3.3 below.

The simplest case for which there is stabilizability is when the dynamical system $\eta(t,x)$ generated by $A+F$, that is the solution of

$$\eta'=A\eta+F(\eta),\qquad \eta(0)=x, \tag{3.1}$$

is "exponentially stable." Indeed, in this case it suffices to take $u=0$ in (1.2). More precisely, we can easily prove the following proposition.

PROPOSITION 3.1. *Assume* (SL) (*respectively*, (SL')). *Let* $h(0)=0$ *and suppose that there exist positive constants* C, R, σ, τ, *and* δ *such that*

$$|\eta(t,x)|\leqq Ce^{-\delta t}|x|^{\tau}\quad \textit{for all } x\in X, \tag{3.2}$$

$$|g(x)|\leqq C|x|^{\sigma}\qquad \textit{for } |x|\leqq R. \tag{3.3}$$

Then, (1.1)-(1.2) *is stable.*

Remark 3.2. A typical assumption that implies (3.2) is that $A+F+\varepsilon$ be dissipative for some $\varepsilon>0$, i.e.,

$$\langle Ax+F(x)+\varepsilon x,x^*\rangle\leqq 0\quad \text{for all } x\in D(A) \text{ and } x^*\in\partial|x|. \tag{3.4}$$

Next, when B is invertible, we can prove a quite general result.

PROPOSITION 3.3. *Assume* (SL) (*respectively*, (SL')) *and let* $B^{-1}\in\mathscr{L}(X;U)$. *Suppose further that there exist positive constants* C, R, *and* σ *such that*

$$|F(x)|\leqq C|x|^{\sigma}\quad \textit{for } |x|\leqq R \ (\textit{respectively}, |F(x)|\leqq C(|x|_{\alpha,p})^{\sigma} \textit{ for } |x|_{\alpha,p}\leqq R), \tag{3.5}$$

$$|g(x)|\leqq C|x|^{\sigma}\quad \textit{for } |x|\leqq R, \tag{3.6}$$

$$|h(u)|\leqq C|u|^{\sigma}\quad \textit{for } |u|\leqq R. \tag{3.7}$$

Then, (1.1)-(1.2) *is stable.*

Proof. We set

$$u(t) = -B^{-1}\{(\omega+1)\, e^{t(A-\omega-1)}x + F(e^{t(A-\omega-1)}x)\}. \tag{3.8}$$

Then

$$|u(t)| \leqq \|B^{-1}\|_{\mathscr{L}(X;U)}\{|\omega+1||x|\, e^{-t} + C|x|\, e^{-\gamma t}\} \quad \text{for } t > \log(|x|/R)/\gamma.$$

So, the corresponding solution of the state equation (1.2) is given by $y(t) = e^{t(A-\omega-1)}x$. In view of (3.5), (3.6), and (3.7), u is an admissible control at x and the proof is complete. □

Now we consider the case when F is "small."

PROPOSITION 3.4. *Assume* (SL) (*respectively* (SL')) *and that there exist positive constants* C, R, *and* σ *such that* (3.5), (3.6), *and* (3.7) *hold. Assume in addition that there exists* $K \in \mathscr{L}(X; U)$ *such that* $A - BK$ *is exponentially stable, i.e., that* (A, B) *is stabilizable by a feedback* K. *There exists* $\varepsilon_0 > 0$ *such that if*

$$\begin{aligned} &|F(x) - F(y)| \leqq \varepsilon_0 |x - y| \quad \textit{for all } x, y \in X \\ &(\textit{respectively}, |F(x) - F(y)| \leqq \varepsilon_0 |x - y|_{\alpha,p} \textit{ for all } x, y \in D_A(\alpha, p)); \end{aligned} \tag{3.9}$$

then (1.1)-(1.2) *is stable.*

Proof. Assume that (3.9) hold for some ε, and let $x \in X$. We will show that, if ε is sufficiently small, then the following control

$$u_x(t) = e^{t(A-BK)}x \tag{3.10}$$

is admissible. Let $N > 0$ and $c > 0$ be such that

$$\|e^{t(A-BK)}\| \leqq N e^{-2ct}, \qquad t \geqq 0 \tag{3.11}$$

and set

$$\begin{aligned} &\|v\|_c = \sup\{e^{ct}|v(t)|;\ t \geqq 0\}, \qquad v \in C([0, \infty[;\ X), \\ &(\text{respectively, } \|v\|_c = \sup\{e^{ct}|v(t)|_{\alpha,p};\ t \geqq 0\}, \qquad v \in C([0, \infty[;\ D_A(\alpha, p)), \end{aligned} \tag{3.12}$$

$$\begin{aligned} &\Sigma = \{v \in C([0, \infty[;\ X);\ \|v\|_c < \infty\} \\ &(\text{respectively, } \Sigma = \{v \in C([0, \infty[;\ D_A(\alpha, p));\ \|v\|_c < \infty\}). \end{aligned} \tag{3.13}$$

Σ, equipped with the norm $\|\ \|_c$, is a Banach space. Now consider the problem

$$z' = (A - BK)z + F(z), \qquad z(0) = x. \tag{3.14}$$

By a fixed point argument we can easily show that if ε is small, then (3.14) has a unique solution in Σ. Since z coincides with the solution of the state equation (1.2) when $u = u_x$, we have obtained the conclusion. □

4. Existence. In this section we prove that the existence of solutions to the Hamilton-Jacobi equation

$$H(B^*DV_\infty(x)) - \langle Ax + F(x), DV_\infty(x)\rangle - g(x) = 0 \tag{4.1}$$

is equivalent to the fact that $(A + F, B, h)$ is g-stabilizable. We will obtain V_∞ as the limit of the generalized viscosity solution to the problem

$$\phi_t(t, x) + H(B^*\nabla\phi(t, x)) - \langle Ax + F(x), \nabla\phi(t, x)\rangle - g(x) = 0, \qquad \phi(0, x) = 0 \tag{4.2}$$

when $t \to +\infty$.

PROPOSITION 4.1. *Assume* (SL) *and suppose that problem* (1.1)-(1.2) *is stable. Let* ϕ *be the generalized viscosity solution to* (4.2) *and let* V_∞ *be given by* (1.4). *Then, for all* $x \in X$ *we have*

$$V_\infty(x) = \lim_{t \uparrow \infty} \phi(t, x). \tag{4.3}$$

Proof. By Proposition 2.5 it follows that $\phi(t, x)$ is increasing in t for any $x \in X$ and $\phi(t, x) \leqq V_\infty(x)$. Thus

$$\phi_\infty(x) = \lim_{t \uparrow \infty} \phi(t, x) \leqq V_\infty(x). \tag{4.4}$$

Now let (u_t, y_t) be such that

$$\phi(t, x) = \int_0^t \{g(y_t(s)) + h(u_t(s))\}\, ds$$

where $u \in L^1(0, t; U)$ and $y_t'(s) = Ay_t(s) + f(y_t(s)) + Bu_t(s)$; $y_t(0) = x$. Then we have

$$V_\infty(x) \geqq \int_0^t h(u_t(s))\, ds \geqq \gamma \|u_t\|^p_{L^p(0,t;H)}. \tag{4.5}$$

Set $\underline{u}_t(s) = u_t(s)$ if $s \in [0, t]$ and $\underline{u}_t(s) = 0$ is $s > t$; since by (4.5) $\{\underline{u}_t\}$ is bounded in $L^p(0, \infty; U)$, there exists

$$t_n \uparrow +\infty \text{ such that } v_n := \underline{u}_{t_n} \to u^* \text{ weakly in } L^p(0, \infty; U); \text{ set } z_n = \underline{y}_{t_n}.$$

Now fix $T > 0$; since e^{tA} is compact for all $t > 0$ (by hypothesis (SL)(ii)) we have that $z_n \to y^*$ in $C([0, T]; X)$, where y^* is the solution of (1.2) with $u = u^*$. Since h is convex it follows that

$$\phi_\infty(x) \geqq \int_0^T \{g(y^*(s)) + h(u^*(s))\}\, ds.$$

But T is arbitrary, so $g(y^*)$ and $h(u^*)$ belong to $L^1(0, \infty; \mathbf{R})$ and

$$\phi_\infty(x) \geqq \int_0^\infty \{g(y^*(s)) + h(u^*(s))\}\, ds \geqq V_\infty(x). \qquad \square$$

Under assumptions (SL′) a similar result can be proved.

PROPOSITION 4.2. *Assume* (SL′) *and suppose that problem* (1.1)-(1.2) *is stable. Let* ϕ *be the generalized viscosity solution to* (4.1) *and* V_∞ *the value function given by* (1.4). *Then, for all* $x \in D_A(\alpha, p)$ *we have*

$$V_\infty(x) = \lim_{t \uparrow \infty} \phi(t, x). \tag{4.6}$$

Proof. The reasoning is similar to the one above. Since F is only defined in $D_A(\alpha, p)$, now we must prove that

$$z_n \to y^* \quad \text{in } C([0, t]; D_A(\alpha, p)). \tag{4.7}$$

From (SL′)(ii) and (1.15) it follows that

$$\frac{d^+}{dt} |z_n(t)|_Z \leqq (a + \omega)|z_n(t)|_Z + |Bv_n(t)| \tag{4.8}$$

where d^+/dt denotes the right derivative. Thus, there exists $C(T)>0$ such that $|z_n(t)|_Z \leqq C(T)$ for every $t\in[0, T]$. We set $\zeta_n = F(z_n)+Bv_n$. Then, from the representation formula

$$z_n(t) = e^{tA}x + \int_0^t e^{(t-s)A}\zeta_n(s)\,ds \tag{4.9}$$

and the fact that v_n is bounded in $L^p(0,\infty; U)$, we conclude that there exists $C_1(T)>0$ such that $|z_n(t)|_{\alpha,p}\leqq C_1(T)$ for every $t\in[0, T]$. Therefore, $\{\zeta_n\}$ is bounded in $L^p(0, T; X)$ and we can find a subsequence, still denoted by $\{\zeta_n\}$, such that $\zeta_n\to\zeta^*$ weakly in $L^p(0, T; X)$. Moreover, $z_n\to y^*$ in $C([0, t]; X)$.

To show (4.7) note that, for all $t, \varepsilon\in]0, T[$,

$$\begin{aligned}|y^*(t)-z_n(t)|_{\alpha,p} &\leqq \int_\varepsilon^T |e^{(t-s)A}\zeta_n(s)|_{\alpha,p}\,ds \\ &\quad + \left(\int_0^\varepsilon \|e^{(t-s)A}\|^{p/(p-1)}_{\mathscr{L}(X,D_A(\alpha,p))}\,ds\right)^{(p-1)/p}\left(\int_0^\varepsilon |\zeta_n(s)|^p\,ds\right)^{1/p}.\end{aligned} \tag{4.10}$$

Also,

$$\|e^{tA}\|_{\mathscr{L}(X,D_A(\alpha,p))}\leqq\frac{\text{const}}{t^\alpha}\quad\forall t>0. \tag{4.11}$$

So, using the fact that e^{tA}, $t>0$, is a compact operator from X into $D_A(\alpha, p)$ and recalling that $\alpha\in]0, 1-1/p[$, we can easily derive (4.7) from (4.10) and (4.11). □

To prove our existence result, we need a lemma.

LEMMA 4.4. *Assume* (SL) (*respectively*, (SL')). *For any* $T>0$ *and* $x\in X$ (*respectively*, $x\in D_A(\alpha, p)$) *we have*

$$\begin{aligned}V_\infty(x) = \inf\Big\{&\int_0^T [g(y(s))+h(u(s))]\,ds + V_\infty(y(T)); \\ &u\in L^1(0, T; U),\ y'(s)+Ay(s)+F(y(s))+Bu(s),\ y(0)=x\Big\}.\end{aligned} \tag{4.12}$$

Proof. Denote by V^* the right-hand side of (4.12). Let u be an admissible control and let y be the corresponding solution of (1.2). Then,

$$\begin{aligned}\int_0^\infty \{g(y(s))+h(u(s))\}\,ds &= \int_0^T \{g(y(s))+h(u(s))\}\,ds \\ &\quad + \int_0^\infty \{g(y(\sigma+T))+h(u(\sigma+T))\}\,d\sigma\end{aligned}$$

whence

$$\int_0^\infty \{g(y(s))+h(u(s))\}\,ds \geqq \int_0^T \{g(y(s))+h(u(s))\}\,ds + V_\infty(y(T)),$$

which implies that $V^*\geqq V_\infty(x)$. We now prove the reverse inequality. Fix $T>0$ and $u\in L^1(0, T; U)$; let $y\in C([0, T]; X)$ be the corresponding solution of (1.2) and (u_T, y_T) be an optimal pair for problem (1.1), (1.2) with $x=y(T)$. Set

$$\underline{u}(s)=u(s)\quad\text{if } 0\leqq s\leqq T,\qquad \underline{u}(s)=u_T(s-T)\quad\text{if } s\geqq T.$$

Since $y_T(0)=y(T)$, we have

$$\underline{y}(s)=y(s)\quad\text{if } 0\leqq s\leqq T,\qquad \underline{y}(s)=y_T(s-T)\quad\text{if } s\geqq T.$$

Then,

$$V_\infty(x) \leqq \int_0^T \{g(y(s))+h(u(s))\}\,ds + \int_T^\infty \{g(y_T(s-T))+h(y_T(s-T))\}\,ds$$
$$= \int_0^T \{g(y(s))+h(u(s))\}\,ds + V_\infty(y(T)),$$

which implies $V_\infty(x) \leqq V^*$. □

The main result of this section is the following theorem.

THEOREM 4.4. *Assume* (SL) (*respectively,* (SL′)) *and suppose that problem* (1.1)-(1.2) *is stable. Then* V_∞ *is a generalized viscosity solution of equation* (4.1). *Moreover, for any* $x \in X$ (*respectively,* $x \in D_A(\alpha, p)$) *there exists an optimal pair* (u^*, y^*) *and the following feedback formula holds*:

$$u^*(t) \in -DH(B^*\nabla^+ V_\infty(y^*(t))), \qquad t \geqq 0. \tag{4.13}$$

Proof. By Lemma 4.3 and by Proposition 2.1 it follows that $V_\infty(x) = W(t, x)$ where W is the generalized viscosity solution of the problem

$$\begin{aligned} -W_t(t,x) + H(B^*DW(t,x)) - \langle Ax+F(x), DW(t,x)\rangle - g(x) = 0, \\ W(T,x) = V_\infty(x). \end{aligned} \tag{4.14}$$

Then, V_∞ is a generalized viscosity solution of (4.1). The existence of an optimal pair (u^*, y^*) was implicitly obtained in the proof of Proposition 4.1 (respectively, Proposition 4.2). Finally, Proposition 2.3 yields the feedback formula (4.13). □

Remark 4.5. From (2.3) we also obtain that, for all $x \in D(A)$,

$$H(B^*p) - \langle Ax+F(x), p\rangle - g(x) \leqq 0 \quad \forall p \in D^+V_\infty(x), \tag{4.15}$$

$$H(B^*p) - \langle Ax+F(x), p\rangle - g(x) \geqq 0 \quad \forall p \in D^-V_\infty(x). \tag{4.16}$$

Remark 4.6—(Maximum principle). Assume (SL) (respectively, (SL′)). From Proposition 2.3 we conclude that, if $x \in X$ (respectively, $x \in D_A(\alpha, p)$) and (u^*, y^*) is an optimal pair at x, then there exists $p^* \in C([0, \infty[; X)$ such that

$$\begin{aligned} &p^{*\prime}(s) + A^*p^*(s) + (DF(y^*(s))^* + Dg(y(^*s))) = 0, \\ &p^*(s) \in D^+V_\infty(y^*(s)), \\ &u^*(s) \in -DH(B^*D^+V_\infty(y^*(s))) \end{aligned} \tag{4.17}$$

for any $s \in [0, T]$.

Remark 4.7—(Feedback dynamical system). Assume (SL) (respectively, (SL′)) and let (u^*, y^*) be an optimal pair at $x \in X$ (respectively, $x \in D_A(\alpha, p)$). Then, by Remark 4.6, y^* is a solution of the closed loop equation

$$y'(t) \in Ay(t) + F(y(t)) - BDH(B^*D^+V_\infty(y(t))), \quad y(0) = x, \quad t \geqq 0. \tag{4.18}$$

Moreover, by Proposition 2.3, there exists $\delta \in]0, 1[$ such that

$$y^* \in C^{1,\delta}(]0, \infty[; X). \tag{4.19}$$

Now, we denote by S_t the dynamical system generated by (4.18), that is,

$$S_t(x) = y(t), \qquad t \geqq 0, \quad x \in X. \tag{4.20}$$

Then, from the Dynamic Programming Principle (4.12) it follows that S_t is a semigroup of nonlinear operators in X.

We remark that no theory is available to directly solve the initial value problem (4.18) except for special situations such as

$$X = U \text{ Hilbert space}, \quad B = 1, \quad h(x) = \tfrac{1}{2}\|x\|^2, \quad V_\infty \text{ convex}.$$

In this case the operator in the right-hand side of (4.18) becomes m-dissipative (see [6]).

Finally we note that

$$g(S_t x) \in L^1(0, \infty; X) \quad \forall x \in X \quad \text{(respectively, } x \in D_A(\alpha, p)\text{)}. \tag{4.21}$$

5. Uniqueness. To make the context of this section clearer to the reader, we recall some known results from linear quadratic control that correspond to the following choice of data:

$$H(x) = \tfrac{1}{2}|x|^2, \quad f(x) = 0, \quad g(x) = \tfrac{1}{2}|Cx|^2, \quad C \in \mathcal{L}(X) \quad x \in X \tag{5.1}$$

where X is a Hilbert space. In this case, setting $V(x) = \frac{1}{2}\langle Px, x\rangle$, (1.3) reduces to the algebraic Riccati equation:

$$A^*P + PA - PBB^*P + C^*C = 0. \tag{5.2}$$

As it is well known, if (A, B) is stabilizable with respect to the observation C, then there exists a minimal positive solution P_∞ of (5.2). Moreover, if the feedback operator

$$L = A - BB^*P_\infty \tag{5.3}$$

is exponentially stable, then P_∞ is unique among the positive solutions of (5.2).

In general, no necessary and sufficient condition for uniqueness of positive solutions is known. A sufficient condition for L to be exponentially stable (which would yield uniqueness), is that C be invertible (more generally that (A, C) be *detectable*; see [25]).

The aim of this section is to generalize the previous results to the general Hamilton-Jacobi equation

$$H(B^*DV(x)) - \langle Ax + F(x), DV(x)\rangle - g(x) = 0. \tag{5.4}$$

Throughout this section we assume either (SL) or (SL′) and that

(5.5) (i) Problem (1.1)-(1.2) is stable;
(ii) $g(0) = 0$, $h(0) = 0$.

By Theorem 4.5 we know that (5.4) has a generalized viscosity solution given by V_∞.

First we remark that V_∞ is minimal.

LEMMA 5.1. *Assume* (SL) (*respectively*, (SL′)) *and* (5.5). *Let V be a nonnegative generalized viscosity solution of* (5.4) *such that* $V(0) = 0$. *Then* $V_\infty(x) \leqq V(x)$, *for all* $x \in X$ (*respectively*, $x \in D_A(\alpha, p)$).

Proof. By Proposition 2.5 it follows that $\phi(t, x) \leqq V(x)$ where ϕ is the solution of (4.2). Then, Propositions 4.1 and 4.2 yield the conclusion. □

Now, to prove uniqueness we must show that V_∞ is maximal. A sufficient condition for maximality is that $B^{-1} \in \mathcal{L}(H; U)$ and the semigroup of nonlinear operators $S_t(x)$, defined in (4.20), be "stable" for any x in X.

LEMMA 5.2. *Assume* (SL) (*respectively*, (SL′)) *and* (5.5). *Suppose that* $B^{-1} \in \mathcal{L}(H; U)$ *and*

$$\forall x \in X \text{ (respectively, } x \in D_A(\alpha, p)) \ \exists r \geqq 1 \text{ such that } t \to S_t(x) \text{ belongs to } L^r(0, \infty; X). \tag{5.6}$$

Then V_∞ is maximal, that is if V is a generalized viscosity solution of (5.4) *such that* $V(0)=0$, *then*

$$V_\infty(x) \geqq V(x) \quad \forall x \in X. \tag{5.7}$$

Proof. Let $x \in X$ (respectively, $x \in D_A(\alpha, p)$) be fixed. Set $y(t)=S_t(x)$. Recalling Proposition 2.4, we have that $D^+V=\partial V$ (see [9]), where ∂V denotes the generalized gradient in the sense of Clarke [11]. So, by (4.19) and Theorem 2.3.10 of [11], we can differentiate the function $t \to V(y(t))$ in the following sense. There exists $q(t) \in D^+V(y(t))$ such that

$$\frac{d}{dt}V(y(t)) = \langle Ay(t)+F(y(t))+Bu(t), q(t)\rangle \geqq -g(y(t))+\langle u(t), B^*q(t)\rangle + H(B^*q(t))$$

$$\geqq -g(y(t))-h(u(t)).$$

The first of the inequalities above follows from (4.15). Hence,

$$V(y(t))+\int_0^t \{g(y(s))+h(u(s))\} \geqq V(x).$$

Since $y \in L^r(0,\infty; X)$, there exists a sequence $\{t_n\} \uparrow \infty$, such that $y(t_n) \to 0$. Thus, by the above inequality, we conclude that $V_\infty(x) \geqq V(x)$ as required. □

Remark 5.3. A sufficient condition that yields (5.6) is the coercivity of g, that is,

$$g(x) \geqq C|x|^r \quad \forall x \in X \tag{5.8}$$

for some constant $C>0$.

From Lemmas 5.1 and 5.2 we deduce the following uniqueness result.

THEOREM 5.4. *Assume* (SL) (*respectively*, (SL′)), (5.5), *and* (5.6). *Then* (5.4) *has a unique generalized viscosity solution that is nonnegative and vanishes at* $x=0$.

6. Application to a semilinear parabolic state equation. Let Ω be a bounded open set of $\mathbf{R}^n$ with smooth boundary $\partial\Omega$. Consider the following optimal control problem:

Minimize

$$J(u,x)=\frac{1}{p}\int_0^\infty dt \int_\Omega \{|y(t,\xi)|^p+|u(t,\xi)|^p\}\, d\xi \tag{6.1}$$

over all controls $u \in L^p([0,\infty[\times\Omega)$, $p>1$, and states y satisfying

$$\frac{\partial y}{\partial t}(t,\xi)=\Delta_\xi y(t,\xi)+\Gamma(y(t,\xi), \nabla_\xi y(t,\xi))+u(t,\xi) \quad \text{in } [0,\infty[\times\Omega, \tag{6.2}$$

$$y(t,\xi)=0 \quad \text{on } [0,\infty[\times\partial\Omega, \tag{6.3}$$

$$y(0,\xi)=x(\xi) \quad \text{on } \Omega \tag{6.4}$$

where $\Gamma(r,s)$ is a real-valued function defined in $\mathbf{R}\times\mathbf{R}^n$ and $x \in L^p(\Omega)$.

To apply the results of §§ 3–5, we proceed to check the assumptions (SL′). Let $X=U=L^p(\Omega)$, A be defined by

$$D(A)=W^{2,p}(\Omega)\cap W_0^{1,p}(\Omega) \quad Az=\Delta_\xi z \quad \forall z \in D(A) \tag{6.5}$$

and let $B=1$. Then A generates an analytic semigroup in $L^p(\Omega)$ by [1] and the embedding of $D(A)$ in $L^p(\Omega)$ is compact in view of the Rellich Theorem. Also, we set

$$g(x)=\frac{1}{p}\int_\Omega |x(\xi)|^p\, d\xi, \qquad h(u)=\frac{1}{p}\int_\Omega |u(\xi)|^p\, d\xi. \tag{6.6}$$

Then, it is well known that $g, h \in C^2(L^p(\Omega))$, provided that

$$p \geqq 2. \tag{6.7}$$

So far, we have shown that assumptions (SL')(i) are satisfied.

Now we check (SL')(ii). For this purpose we define

$$Z = C(\bar{\Omega})$$

and note that by the results of [23] the part of A in Z, A_Z, generates an analytic semigroup. This semigroup is also contracting in view of the maximum principle and so (SL')(ii) holds with $\mu = 0$. Next, to verify (SL')(iii), recall the following well-known characterization of the interpolation spaces $D_A(\alpha, p)$ (see, for instance, [24]):

$$D_A(\alpha, p) = \begin{cases} \{f \in W^{2\alpha,p}(\Omega); f_{|\partial\Omega} = 0\} & \text{if } \alpha \in \left]\frac{1}{2p}, 1\right[, \\ W^{2\alpha,p}(\Omega) & \text{if } \alpha \in \left]0, \frac{1}{2p}\right[. \end{cases}$$

By the Sobolev Embedding Theorem,

$$D_A(\alpha, p) \subset C(\bar{\Omega}) = z \quad \text{if } \alpha > \frac{n}{2p}. \tag{6.8}$$

Note that the constraint in (6.8) is compatible with the requirement $\alpha \in]0, 1 - 1/p[$ if

$$p > \frac{n+2}{2}. \tag{6.9}$$

Let $F(x) = \Gamma(x, \nabla_\xi x)$ and assume

$$\Gamma \in C^2(\mathbf{R} \times \mathbf{R}^n). \tag{6.10}$$

From the Sobolev Embedding Theorem it follows that

$$\alpha > \frac{n+p}{2p} \Rightarrow W^{2\alpha,p}(\Omega) \subset C^1(\bar{\Omega}), \tag{6.11}$$

which in turn implies that F fulfills (1.14). Note again that the constraint in (6.11) is compatible with the requirement $\alpha \in]0, 1 - 1/p[$ if

$$p > n + 2. \tag{6.12}$$

We will now show that the condition

$$r\Gamma(r, 0) \leqq ar^2 \quad \text{for all } r \in \mathbf{R} \text{ and some } a \in \mathbf{R} \tag{6.13}$$

implies (1.15). The argument is known; nevertheless, we recall it for the reader's convenience. First, let

(6.14) $z \in C^1(\bar{\Omega})$ be such that $|z|$ has a unique maximum point, say $\xi_0 \in \Omega$.

Then, we can easily show that $\partial|z| = \{z^*\}$, where

$$z^* = \begin{cases} \delta_{\xi_0} & \text{if } z(\xi_0) = |z|_Z, \\ -\delta_{\xi_0} & \text{if } z(\xi_0) = -|z|_Z, \end{cases}$$

and δ denotes the Dirac measure. Thus,

$$\langle F(z), z^* \rangle = \begin{cases} \Gamma(|z|_Z, 0) & \text{if } z(\xi_0) = |z|_Z, \\ -\Gamma(-|z|_Z, 0) & \text{if } z(\xi_0) = -|z|_Z. \end{cases} \tag{6.15}$$

From (6.13) and (6.15) we get

$$\langle F(z), z^*\rangle \leqq a|z|_Z \tag{6.16}$$

for all z satisfying (6.14). On the other hand, it is well known (see, for instance, [17, Lemma II-7-1]) that $z \in Z$ satisfies (6.16) if and only if

$$|z| \leqq |z+\lambda(F(z)-az)| \quad \forall \lambda > 0. \tag{6.17}$$

Since the set of functions z satisfying (6.14) is dense in Z, the proof of (1.15) is complete. Finally, if we assume that

$$|\Gamma(r, s)| \leqq \beta(|r|)+\rho(|r|)|s| \tag{6.18}$$

where $\beta, \rho : [0, \infty[\to [0, \infty[$ are continuous functions, then (1.16) easily follows. Therefore, assumptions (SL′) are fulfilled if

(6.19) $(n+p)/2p < \alpha < 1-1/p$, $p > n+2$, and (6.10), (6.13), and (6.18) hold.

Our next goal is to show that $(A+F, B, h)$ is g-stabilizable. This will be given by Proposition 3.3 if we assume that the function β in (6.18) satisfies

$$\beta(r) \leqq Cr^\sigma \quad \forall r \in [0, R] \tag{6.20}$$

for some constants $C, R \geqq 0$.

Now, Theorem 5.4 yields the following theorem.

THEOREM 6.1. *Assume* (6.19) *and* (6.20). *Then the Hamilton-Jacobi equation*

$$(p-1)|DV(x)|_{X^*}^{p'} - p\langle \Delta_\xi x + \Gamma(x, \nabla_\xi x), DV(x)\rangle - |x|_X^p = 0, \qquad p' = \frac{p}{p-1} \tag{6.21}$$

has a unique generalized viscosity solution $V_\infty \geqq 0$ *such that* $V_\infty(0) = 0$. V_∞ *is the value function of the control problem* (6.1)-(6.4). *Moreover, for any* $x \in D_A(\alpha, p)$ *there exists an optimal pair* (u^*, y^*) *at* x *and the following feedback formula holds*:

$$u^*(t) \in |D^+V(y^*(t))|^{p'-2}D^+V(y^*(t)). \tag{6.22}$$

REFERENCES

[1] S. AGMON, *On the eigenfunctions and the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119-147.

[2] H. AMANN, *Feedback stabilization of linear and semilinear parabolic systems*, in Proc. Trends in Semigroup Theory and Applications, Trieste, September 28-October 2, 1987, Lecture Notes in Pure and Applied Mathematics, Marcel Dekker, New York, 1988.

[3] V. BARBU, *On convex control problems on infinite intervals*, J. Math. Anal. Appl., 65 (1978), pp. 687-702.

[4] ———, *Hamilton-Jacobi equations and non-linear control problems*, J. Math. Anal. Appl., 120 (1986), pp. 494-509.

[5] ———, *Optimal Control of Variational Inequalities*, Pitman, Boston, 1984.

[6] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, Boston, 1982.

[7] ———, *Hamilton-Jacobi equations in Hilbert spaces. Variational and semigroup approach*, Ann. Mat. Pura Appl., 142 (1985), pp. 303-349.

[8] V. BARBU, E. N. BARRON, AND R. JENSEN, *The necessary conditions for optimal control in Hilbert spaces*, Math. Anal. Appl., 133 (1988), pp. 151-162.

[9] P. CANNARSA, *Regularity properties of solutions to Hamilton-Jacobi equations in infinite dimensions and nonlinear optimal control*, Differential and Integral Equations, to appear.

[10] P. CANNARSA AND G. DA PRATO, *Some results on nonlinear optimal control problems and Hamilton-Jacobi equations in infinite dimensions*, preprint.

[11] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[12] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 183-186.

[13] M. G. CRANDALL AND P.-L. LIONS, *Hamilton-Jacobi equations in infinite dimensions. I. Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379-396.

[14] ———, *Hamilton-Jacobi equations in infinite dimensions. II. Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368-405.

[15] ———, *Hamilton-Jacobi equations in infinite dimensions.* III. J. Funct. Anal., 68 (1986), pp. 214-247.

[16] ———, *Solutions de viscosité pour les équations de Hamilton-Jacobi en dimension infinie intervénant dans le contrôle optimal de problèmes d'évolution*, C. R. Acad. Sci. Paris, 305 (1987), pp. 233-236.

[17] G. DA PRATO, *Applications croissantes et équations d'évolutions dans les espaces de Banach*, Academic Press, London, 1976.

[18] I. LASIECKA, *Stabilization of hyperbolic and parabolic systems with nonlinearly perturbed boundary conditions*, J. Differential Equations, to appear.

[19] J.-L. LIONS AND J. PEETRE, *Sur une classe d'espaces d'interpolation*, Inst. Hautes Etudes Sci. Publ. Math., 19 (1964), pp. 5-68.

[20] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[21] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[22] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1981), pp. 25-52.

[23] H. B. STEWART, *Generation of analytic semigroup by strongly elliptic operators under general boundary conditions*, Trans. Amer. Math. Soc., 259 (1980), pp. 299-310.

[24] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

[25] J. ZABCZYK, *Remarks on the algebraic Riccati equations in Hilbert spaces*, Appl. Math. Optim., 2 (1976), pp. 251-258.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 4, pp. 876-907, July 1989

012

REGULARITY OF THE VALUE FUNCTION FOR A TWO-DIMENSIONAL SINGULAR STOCHASTIC CONTROL PROBLEM*

H. METE SONER† AND STEVEN E. SHREVE‡

Abstract. It is desired to control a two-dimensional Brownian motion by adding a (possibly singularly) continuous process to it so as to minimize an expected infinite-horizon discounted running cost. The Hamilton-Jacobi-Bellman characterization of the value function V is a variational inequality which has a unique *twice* continuously differentiable solution. The optimal control process is constructed by solving the Skorokhod problem of reflecting the two-dimensional Brownian motion along a free boundary in the $-\nabla V$ direction.

Key words. singular stochastic control, variational inequality, free boundary problem, Skorokhod problem

AMS(MOS) subject classifications. 93E20, 35R35

1. Introduction.

We study regularity of the solution of the variational inequality associated with a two-dimensional singular stochastic control problem with a convex running cost. The solution u of this variational inequality, which is the value function for the control problem, is shown to be of class C^2. We also study the regularity of the free boundary in $\mathbb{R}^2$ which divides the region where u satisfies a second-order elliptic equation from the region where it does not. The free boundary is shown to be smooth, and this fact is instrumental in our construction of the optimal process for the stochastic control problem.

Previous work on the regularity of the value function in singular stochastic control has focused on one-dimensional problems. Beneš, Shepp, and Witsenhausen (1980) suggested that the value function for these problems should be of class C^2 and used this so-called "principle of smooth fit" to determine some otherwise free parameters that arose in the solution of their problems. It has been used in the same way by Harrison (1985), Harrison and Taylor (1978), Harrison and Taksar (1983), Karatzas (1981), (1983), Lehoczky and Shreve (1986), Shreve, Lehoczky, and Gaver (1984), and Taksar (1985). (But see Menaldi and Robin (1983), Chow, Menaldi, and Robin (1985), and Sun (1987) for a variational inequality approach to singular control that does not use the principle of smooth fit.) An important question is whether the principle of smooth fit can be expected to apply to multidimensional singular control problems, or is it strictly a one-dimensional phenomenon. Karatzas and Shreve (1986) suggested that it might apply in higher dimensions. These authors studied the singular control of a one-dimensional Brownian motion under a constraint on the total variation of the control process (a "finite-fuel" constraint). The fuel remaining constitutes a second state variable, and the value function for this problem was found to be of class C^2 jointly in both state variables. One should observe, however, that the second state variable in this problem is not a diffusion; indeed, the fuel remaining is constant until control is exercised, at which time it decreases an amount equal to the displacement caused by the control.

* Received by the editors August 17, 1988; accepted for publication (in revised form) December 15, 1988.

† Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This work was supported by National Science Foundation grant DMS-87-02537.

‡ Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This work was supported by Air Force Office of Scientific Research grants AFOSR-85-0360 and AFOSR-89-0075.

This paper concerns the control of a two-dimensional Brownian motion, and control can cause displacement in any direction. Thus, the discovery of a C^2 value function provides strong support for belief in a widely applicable principle of smooth fit. Nevertheless, the argument of this paper depends heavily on the fact that only two dimensions are involved (see Remark 6.2), and we have not found a way to obtain a similar result in higher dimensions.

This paper is organized as follows. Section 2 defines the underlying stochastic control problem, and § 3 relates it to a free boundary problem, the so-called Hamilton-Jacobi-Bellman (HJB) equation. Section 4 constructs a $C^{1,1}$, nonnegative convex solution u to the HJB equation and proves its uniqueness. Sections 5–10 upgrade the regularity of u to C^2. The key idea here is to use the gradient flow of u to change to a more convenient pair of coordinates. This is a generalization of the device used by many authors in one-dimensional problems of differentiating the Bellman equation so as to obtain a more standard free boundary problem. In § 11 the free boundary is shown to be of class $C^{2,\alpha}$ for any $\alpha \in (0, 1)$. In § 12 we return to the stochastic control problem, which now reduces to the Skorokhod problem of finding a Brownian motion reflected along the free boundary in the $-\nabla u$ direction. The established regularity of u and the free boundary allow us to assert the existence and uniqueness of a solution to the Skorokhod problem and finally complete the proof, begun in § 3, that u is the value function for the stochastic control problem of § 2.

2. The singular stochastic control problem. Let $\{W_t, \mathcal{F}_t; 0 \leqq t < \infty\}$ be a standard, two-dimensional Brownian motion defined on a complete probability space $(\Omega, \mathcal{F}, P)$, and let $\{\mathcal{F}_t\}$ be the augmentation of the filtration generated by W (see Karatzas and Shreve (1987, p. 89)). The *state process* for our control problem is

$$X_t \triangleq x + \sqrt{2}\, W_t + \int_0^t N_s \, d\zeta_s, \qquad 0 \leqq t < \infty, \tag{2.1}$$

where $x \in \mathbb{R}^2$ is the initial condition and the *control process pair* $\{(N_t, \zeta_t); 0 \leqq t < \infty\}$ is $\{\mathcal{F}_t\}$-adapted and satisfies the conditions:

$$|N_t| = 1, \quad \forall 0 \leqq t < \infty \quad \text{a.s.}, \tag{2.2}$$

where $|\cdot|$ denotes the Euclidean norm, and

$$\zeta \text{ is nondecreasing, left-continuous, and } \zeta_0 = 0 \quad \text{a.s.}. \tag{2.3}$$

The process N gives the direction and ζ gives the intensity of the "push" applied by the controller to the state X.

Given control processes N and ζ, we define the corresponding cost

$$V_{N,\zeta}(x) \triangleq E^x \int_0^\infty e^{-t}[h(X_t)\, dt + d\zeta_t], \tag{2.4}$$

where $h: R^2 \to \mathbb{R}$ is a strictly convex function satisfying, for appropriate positive constants C_0, c_0, and q:

$$h \in C^{2,1}_{loc}(\mathbb{R}^2), \tag{2.5}$$

$$0 \leqq h(x) \leqq C_0(1 + |x|^q) \quad \forall x \in \mathbb{R}^2, \tag{2.6}$$

$$|\nabla h(x)| \leqq C_0(1 + h(x)) \quad \forall x \in \mathbb{R}^2, \tag{2.7}$$

$$c_0 |y|^2 \leqq D^2 h(x) y \cdot y \leqq C_0 |y|^2 (1 + h(x)) \quad \forall x \in \mathbb{R}^2, y \in \mathbb{R}^2. \tag{2.8}$$

Without loss of generality, we also assume that

$$0 = h(0) \leqq h(x) \quad \forall x \in \mathbb{R}^2. \tag{2.9}$$

For $x \in \mathbb{R}^2$, we define the *value function*

$$V(x) \triangleq \inf_{N,\zeta} V_{N,\zeta}(x). \tag{2.10}$$

3. The Hamilton–Jacobi–Bellman equation. We shall show that the value function V of (2.10) is characterized by the *Hamilton-Jacobi-Bellman* (HJB) equation

$$\max\{u - \Delta u - h, |\nabla u|^2 - 1\} = 0. \tag{3.1}$$

The following theorem gives a partial description of the relationship between V and the HJB equation. More definitive results are proved in § 12.

THEOREM 3.1. *Let $u: \mathbb{R}^2 \to \mathbb{R}$ be a convex, C^2 solution of* (3.1). *Then $u \leqq V$. For a given $x \in \mathbb{R}^2$, suppose there exists a control process pair (N, ζ) such that $V_{N,\zeta}(x) < \infty$ and the corresponding state process* (2.1) *satisfies*

$$u(X_t) - \Delta u(X_t) - h(X_t) = 0 \quad \forall t \in (0, \infty), \quad \text{a.s.}, \tag{3.2}$$

$$\int_0^t 1_{\{N_s = -\nabla u(X_s)\}}\, d\zeta_s = \zeta_t \quad \forall t \in [0, \infty), \quad \text{a.s.}, \tag{3.3}$$

$$u(X_t) - u(X_{t+}) = \zeta_{t+} - \zeta_t \quad \forall t \in [0, \infty), \quad \text{a.s.} \tag{3.4}$$

Then

$$u(x) = V(x) = V_{N,\zeta}(x),$$

i.e., (N, ζ) is optimal at x.

Proof. Let $x \in \mathbb{R}^2$ and any control process pair (N, ζ) be given. Applying Itô's rule for semimartingales (Meyer (1976, pp. 278, 301)) to $e^{-t}u(X_t)$, adjusting the result to account for the fact that ζ is left-continuous rather than right-continuous, and observing that $|\nabla u| \leqq 1$ so $E\int_0^t e^{-s}\nabla u(X_s)\, dW_s = 0$, we obtain for $t \geqq 0$:

$$\begin{aligned} u(x) = {} & E\, e^{-t}u(X_t) + E\int_0^t e^{-s}[u(X_s) - \Delta u(X_s) - h(X_s)]\, ds \\ & + E\int_0^t e^{-s}h(X_s)\, ds + E\int_0^t [-e^{-s}\nabla u(X_s)\cdot N_s]\, d\zeta_s \\ & + E\sum_{0 \leqq s < t} e^{-s}[u(X_s) - u(X_{s+}) + \nabla u(X_s)\cdot N_s(\zeta_{s+} - \zeta s)]. \end{aligned} \tag{3.5}$$

The second and fifth terms on the right-hand side of (3.5) are nonpositive because of (3.1) and the convexity of u, respectively. Because $|\nabla u| \leqq 1$, the fourth term is dominated by $E\int_0^t e^{-s}\, d\zeta_s$, and thus we have

$$u(x) \leqq E\, e^{-t}u(X_t) + E\int_0^t e^{-s}[h(X_s)\, ds + d\zeta_s]. \tag{3.6}$$

We wish to let $t \to \infty$ in (3.6) to obtain

$$u(x) \leqq E\int_0^\infty e^{-s}[h(X_s)\, ds + d\zeta_s] = V_{N,\zeta}(x). \tag{3.7}$$

Assume $E\int_0^\infty e^{-s}h(X_s)<\infty$, for otherwise (3.7) is obviously true. This implies that

$$\underline{\lim}_{t\to\infty} E\,e^{-t}h(X_t)=0.$$

Now (2.8), (2.9), and the inequality $|\nabla u|\leq 1$ (from (3.1)) imply that for all $y\in\mathbb{R}^2$,

$$u(y)\leqq u(0)+|y|\leqq u(0)+1+|y|^2\leqq u(0)+1+\frac{2}{c_0}h(y), \tag{3.8}$$

so

$$\underline{\lim}_{t\to\infty} E\,e^{-t}u(X_t)=0.$$

We may therefore pass to the limit in (3.6) along a sequence $\{t_n\}_{n=1}^\infty$ such that $E\,e^{-t_n}u(X_{t_n})\to 0$ as $t_n\to\infty$, and (3.7) follows. Since (N,ζ) is an arbitrary control process pair, we have $u(x)\leqq V(x)$.

If (3.2)–(3.4) are satisfied, then the second and fifth terms on the right-hand side of (3.5) are zero, and the fourth term is $E\int_0^t e^{-s}\,d\zeta_s$. It follows that equality holds in (3.6), and hence also in (3.7), i.e.,

$$u(x)\leqq V(x)\leqq V_{(N,\zeta)}(x)=u(x). \qquad \square$$

Remark 3.2. Equation (3.1) is similar but not equivalent to a problem arising in elastic-plastic torsion (Ting (1966), (1967), Duvant and Lanchon (1967), Brezis and Sibony (1971)). The elastic-plastic problem is posed on a bounded domain $\Omega\subset\mathbb{R}^n$, and is to minimize

$$J(v)\triangleq\int_\Omega \frac{1}{2}|\nabla v|^2-vh$$

over $K\triangleq\{v\in H_0^1(\Omega);\ \|\nabla v\|_\infty\leqq 1\}$. Equivalently, one seeks $u\in K$ satisfying

$$\int h(v-u)-\int \nabla u\cdot(\nabla v-\nabla u)\leqq 0 \quad \forall v\in K.$$

If u solves the elastic-plastic torsion problem, then

$$(\Delta u+h)(|\nabla u|-1)=0,$$

but $\Delta u+h$ may be negative. In the special case that h is a nonnegative constant function, a solution to the elastic-plastic problem also satisfies a variational inequality like (3.1) (see Evans (1979, § 6), but such an h is not interesting in the control problem.

4. Solution of the Hamilton–Jacobi–Bellman equation. The existence of a $W^{2,\infty}_{loc}$ solution to the HJB equation (3.1) follows from a modification of Evans (1979) (see also Ishii and Koike (1983)), who treated a bounded domain and general h and space dimension. We need to refer to this construction in the next section, so we provide it here.

Let $\beta:\mathbb{R}\to\mathbb{R}$ be a C^∞ function satisfying

$$\beta(r)=0 \quad \forall r\in(-\infty,0], \tag{4.1i}$$

$$\beta(r)>0 \quad \forall r\in(0,\infty), \tag{4.1ii}$$

$$\beta(r)=r-1 \quad \forall r\in[2,\infty), \tag{4.1iii}$$

$$\beta'(r)\geqq 0,\quad \beta''(r)\geqq 0 \quad \forall r\in\mathbb{R}. \tag{4.1iv}$$

For each $\varepsilon > 0$, we form the penalization function

$$\beta_\varepsilon(r) \triangleq \beta\left(\frac{r-1}{\varepsilon}\right) \quad \forall r \in \mathbb{R}, \tag{4.2}$$

and we consider the *penalized equation*

$$u^\varepsilon - \Delta u^\varepsilon + \beta_\varepsilon(|\nabla u^\varepsilon|^2) = h. \tag{4.3}$$

The following lemma is proved in the appendix.

LEMMA 4.1. *For every $\varepsilon \in (0,1)$, there exists a nonnegative, convex, C^2solution u^ε* to (4.3). *There exist positive constants C_1, C_2, and p, indepenent of ε, such that for all $\varepsilon \in (0,1)$, for all $x \in \mathbb{R}^2$:*

$$0 \leqq u^\varepsilon(x) \leqq C_1(1+|x|^p), \tag{4.4}$$

$$|\nabla u^\varepsilon(x)| \leqq C_1(1+|x|^p), \tag{4.5}$$

and for every $y \in \mathbb{R}^2$,

$$0 \leqq D^2 u^\varepsilon(x) y \cdot y \leqq C_2 |y|^2 (1+u^\varepsilon(x)). \tag{4.6}$$

DEFINITION 4.2. We define a norm on the vector space of 2×2 matrices by

$$\|A\| \triangleq \sqrt{\text{trace}\,(AA^T)}.$$

If A is symmetric with eigenvalues λ_1 and λ_2, then

$$\|A\| = \sqrt{\lambda_1^2 + \lambda_2^2}. \tag{4.7}$$

THEOREM 4.3. *The* HJB *equation* (3.1) *has a nonnegative, convex solution $u \in W^{2,\infty}_{loc}$ satisfying*

$$\|D^2 u(x)\| \leqq C_3(1+|x|^m), \quad \text{a.e. } x \in \mathbb{R}^2, \tag{4.8}$$

for some $C_3 > 0$ and $m \in \mathbb{N}$.

Proof. Because $D^2 u^\varepsilon$ is locally bounded uniformly in $\varepsilon \in (0,1)$, we may choose a decreasing sequence $\{\varepsilon_n\}_{n=1}^\infty$ with limit zero such that $\{u^{\varepsilon_n}\}_{n=1}^\infty$ and $\{\nabla u^{\varepsilon_n}\}_{n=1}^\infty$ converge uniformly on compact sets, and $\{D^2 u^{\varepsilon_n}\}_{n=1}^\infty$ converges in the L^∞_{loc}-weak* topology. Define $u = \lim_{n\to\infty} u^{\varepsilon_n}$, so that $\nabla u = \lim_{n\to\infty} \nabla u^{\varepsilon_n}$ and the weak* limit of $\{D^2 u^{\varepsilon_n}\}_{n=1}^\infty$ is $D^2 u$. Passage to the limit in (4.3) gives (3.1). □

LEMMA 4.4. *Let $u \in W^{2,\infty}_{loc}$ be a nonnegative, convex solution to the* HJB *equation* (3.1), *and define*

$$\mathscr{C} \triangleq \{x \in \mathbb{R}^2; |\nabla u(x)|^2 < 1\}. \tag{4.9}$$

Then for every unit vector ν,

$$u_{\nu\nu} \triangleq (D^2 u)\nu \cdot \nu > 0 \quad \text{on } \mathscr{C}. \tag{4.10}$$

$\mathscr{C}$ is bounded, and u attains its unique minimum over $\mathbb{R}^2$ inside $\mathscr{C}$.

Proof. We have

$$u - \Delta u = h \quad \text{on } \mathscr{C}, \tag{4.11}$$

and $h \in C^{2,1}_{loc}(\mathbb{R}^2)$, so $u \in C^{4,\alpha}(\mathscr{C})$ for all $\alpha \in (0,1)$. Differentiating (4.11), we obtain

$$u_{\nu\nu} - \Delta u_{\nu\nu} = h_{\nu\nu} \quad \text{on } \mathscr{C}.$$

and since $h_{\nu\nu} > 0$, relation (4.10) holds. Equation (4.11) also implies that $u \geqq h$ on $\mathscr{C}$, and since $|\nabla u| \leqq 1$ on $\mathbb{R}^2$ but h grows at least quadratically (see (2.8)), $\mathscr{C}$ must be bounded.

Let $\delta \in (0, \frac{1}{2})$ be given, and choose $x^\delta \in \mathbb{R}^2$ such that

$$u(x^\delta) \leqq u(x) + \delta \quad \forall x \in \mathbb{R}^2.$$

Define

$$\psi_\delta(x) \triangleq u(x) + \delta|x - x^\delta|^2 \quad \forall x \in \mathbb{R}^2,$$

and note that ψ_δ attains its minimum over $\mathbb{R}^2$ at some point y^δ. In particular,

$$0 = \nabla \psi_\delta(y^\delta) = \nabla u(y^\delta) + 2\delta(y^\delta - x^\delta). \tag{4.12}$$

But also

$$u(y^\delta) + \delta|y^\delta - x^\delta|^2 = \psi_\delta(y^\delta) \leqq \psi_\delta(x^\delta) = u(x^\delta) \leqq u(y^\delta) + \delta.$$

It follows that $|y^\delta - x^\delta| \leqq 1$, and returning to (4.12), we see that $|\nabla u(y^\delta)| \leqq 2\delta < 1$. Therefore, $y^\delta \in \mathscr{C}$ for all $\delta \in (0, \frac{1}{2})$, and the sequence $\{y^{1/n}\}_{n=3}^\infty$ accumulates at some $y^0 \in \bar{\mathscr{C}}$. From (4.12) we have $\nabla u(y^0) = 0$, so $y^0 \in \mathscr{C}$, and the convexity of u on $\mathbb{R}^2$ implies that u attains its minimum at y_0. This minimum is unique because of (4.10). □

THEOREM 4.5. *There is only one nonnegative, convex solution $u \in W^{2,\infty}_{loc}$ to the* HJB *equation* (3.1).

Proof. *Let* u_1 and u_2 be two nonnegative, convex solutions to (3.1), and let y^0 be the point where u_2 attains its minimum. Given $\delta > 0$, define

$$\varphi_\delta(x) \triangleq u_1(x) - u_2(x) - \delta|x - y^0|^2 \quad \forall x \in \mathbb{R}^2.$$

The function φ_δ attains its maximum at some $x^\delta \in \mathbb{R}^2$, and $0 = \nabla \varphi_\delta(x^\delta) = \nabla u_1(x^\delta) - \nabla u_2(x^\delta) - 2\delta(x^\delta - y^0)$. Consequently,

$$1 \geqq |\nabla u_1(x^\delta)|^2 = |\nabla u_2(x^\delta)|^2 + 4\delta^2|x^\delta - y^0|^2 + 4\delta \nabla u_2(x^\delta) \cdot (x^\delta - y^0).$$

Because u_2 is convex, $\nabla u_2(x^\delta) \cdot (x^\delta - y^0) \geqq 0$, so either $|\nabla u_2(x^\delta)|^2 < 1$ or $x^\delta = y^0$. This last equality would imply that $\nabla u_2(x^\delta) = 0$, so in any event, $|\nabla u_2(x^\delta)|^2 < 1$. From (3.1) we have

$$\Delta u_2(x^\delta) = u_2(x^\delta) - h(x^\delta).$$

Because φ attains its maximum at x^δ, we have from the Bony maximum principle (Bony (1967), Lions (1983))

$$\begin{aligned} 0 &\geqq \liminf_{x \to x^\delta} \operatorname{ess}\, \Delta\, \varphi_\delta(x) \\ &= \liminf_{x \to x^\delta} \operatorname{ess}\, [\Delta u_1(x) - \Delta u_2(x) - 4\delta] \\ &\geqq u_1(x^\delta) - u_2(x^\delta) - 4\delta. \end{aligned}$$

It follows that for all $x \in \mathbb{R}^2$,

$$\begin{aligned} u_1(x) - u_2(x) &= \varphi_\delta(x) + \delta|x - y^0|^2 \\ &\leqq \varphi_\delta(x^\delta) + \delta|x - y^0|^2 \\ &\leqq \delta(4 + |x - y^0|^2). \end{aligned}$$

Letting $\delta \downarrow 0$, we obtain $u_1 \leq u_2$. The reverse inequality is proved by interchanging u_1 and u_2. □

Remark 4.6. Throughout the remainder of the paper, u will denote the unique nonnegative, convex solution in $W^{2,\infty}_{loc}$ to (3.1). The set $\mathscr{C}$ will be given by (4.9), and $y^0 \in \mathscr{C}$ will denote the unique minimizer of u. We shall prove that $u \in C^{2,\alpha}_{loc}(\mathbb{R}^2)$ for all $\alpha \in (0,1)$ (Theorem 10.3), $\partial\mathscr{C}$ is of class $C^{2,\alpha}$ for all $\alpha \in (0,1)$ (Corollary 11.3), and $n(x) \cdot \nabla u(x) \geqq \sigma$ for all $x \in \partial\mathscr{C}$, where $n(x)$ is the outward normal to $\mathscr{C}$ at x and σ is a positive constant (Lemma 12.2).

5. An obstacle problem. Let us return to the construction of u in the proof of Theorem 4.3 as the limit of a sequence of functions $\{u^{\varepsilon_n}\}_{n=1}^{\infty}$, where each u^{ε_n} satisfies (4.3). Define $w^{\varepsilon_n} \triangleq |\nabla u^{\varepsilon_n}|^2$ and compute the product of ∇u^{ε_n} with the gradient of both sides of (4.3) to obtain

$$w^{\varepsilon_n} - \tfrac{1}{2}\Delta w^{\varepsilon_n} + 2\beta'_{\varepsilon_n}(w^{\varepsilon_n})(D^2 u^{\varepsilon_n})\nabla u^{\varepsilon_n} \cdot \nabla u^{\varepsilon_n} = H^{\varepsilon_n} \tag{5.1}$$

where

$$H^{\varepsilon_n} \triangleq \nabla h \cdot \nabla u^{\varepsilon_n} - \|D^2 u^{\varepsilon_n}\|^2.$$

Along a subsequence, which we also call $\{\varepsilon_n\}_{n=1}^{\infty}$, $\{H^{\varepsilon_n}\}_{n=1}^{\infty}$ converges to

$$\bar{H} \triangleq \nabla h \cdot \nabla u - \chi, \tag{5.2}$$

where χ is the limit of $\|D^2 u^{\varepsilon_n}\|^2$ in the weak* topology on L^{∞}_{loc}. We will show that

$$w = |\nabla u|^2 = \lim_{n \to \infty} |\nabla u^{\varepsilon_n}|^2$$

solves an obstacle problem involving $\bar{H}$, and we will then obtain $W^{2,p}_{loc}$ regularity for w by invoking the theory of variational inequalities.

For $r > 0$ chosen so that $B_r(0) \triangleq \{x \in \mathbb{R}^2; |x| < r\}$ contains $\mathscr{C}$, define

$$K_r \triangleq \{v \in W^{1,2}(B_r);\ 0 \leqq v \leqq 1 \text{ on } B_r \text{ and } v - 1 \in W^{1,2}_0(B_r)\}.$$

We pose the problem of finding $\varphi \in K_r$ such that

$$\frac{1}{2}\int_{B_r(0)} \nabla\varphi \cdot (\nabla v - \nabla\varphi) \geqq \int_{B_r(0)} (\bar{H} - w)(v - \varphi) \quad \forall v \in K_r. \tag{5.3}$$

LEMMA 5.1. *The function* $w = |\nabla u|^2$ *solves* (5.3).

Proof. Let $v \in K_r$ be given. From (5.1) we have

$$\begin{aligned} &\int_{B_r(0)} \left(w^{\varepsilon_n} - \frac{1}{2}\Delta w^{\varepsilon_n} - H^{\varepsilon_n} \right)(v - w^{\varepsilon_n}) \\ &\qquad = -\int_{B_r(0)} 2\beta'_{\varepsilon_n}(w^{\varepsilon_n})(D^2 u^{\varepsilon_n})\nabla u^{\varepsilon_n} \cdot \nabla u^{\varepsilon_n}(v - w^{\varepsilon_n}). \end{aligned} \tag{5.4}$$

The function u^{ε_n} is convex, $\beta'_{\varepsilon_n}(w^{\varepsilon_n}) = 0$ whenever $w^{\varepsilon_n} \leqq 1$, and $v - w^{\varepsilon_n} < 0$ whenever $w^{\varepsilon_n} > 1$. Therefore, the right-hand side of (5.4) is nonnegative, and integration by parts yields

$$\begin{aligned} &-\frac{1}{2}\int_{\partial B_r(0)} (v - w^{\varepsilon_n})\nabla w^{\varepsilon_n} \cdot n + \frac{1}{2}\int_{B_r(0)} \nabla w^{\varepsilon_n} \cdot (\nabla v - \nabla w^{\varepsilon_n}) \\ &\qquad \geqq \int_{B_r(0)} (H^{\varepsilon_n} - w^{\varepsilon_n})(v - w^{\varepsilon_n}), \end{aligned} \tag{5.5}$$

where n is the outward normal on $\partial B_r(0)$. Now $w^{\varepsilon_n} \to v$ uniformly on $\partial B_r(0)$, $w^{\varepsilon_n} \to w$ uniformly on $B_r(0)$, and $H^{\varepsilon_n} \to \bar{H}$, $\nabla w^{\varepsilon_n} \to \nabla w$, both the latter convergences being weak*

in $L^\infty(B_r(0))$. Because the weak* limit of $|\nabla w^{\varepsilon_n}|^2$ dominates $|\nabla w|^2$, we may pass to the limit in (5.5) to obtain

$$\frac{1}{2}\int_{B_r(0)} \nabla w\cdot(\nabla v-\nabla w)\geqq \int_{B_r(0)} (\bar H - w)(v-w) \quad \forall v\in K_r. \tag{5.6}$$ □

THEOREM 5.2. *For every* $p\in(1,\infty)$, $w\triangleq|\nabla u|^2\in W^{2,p}_{loc}$.

Proof. This is a classical result. See, for example, Lemma 5.1 and Theorem 3.11, p. 29 of Chipot (1984). □

COROLLARY 5.3. *We have* $w\in C^{1,\alpha}(\mathbb{R}^2)$ *for any* $\alpha\in(0,1)$.

Proof. This follows from Sobolev imbedding (Gilbarg and Trudinger (1983, Thm. 7.17, p. 163)). □

Remark 5.4. Integration by parts allows us to rewrite (5.6) as

$$\int_{B_r(0)} (w-\tfrac{1}{2}\Delta w-\bar H)(v-w)\geqq 0 \quad \forall v\in K_r,$$

for all sufficiently large r, and so

$$\max\{w-\tfrac{1}{2}\Delta w-\bar H,\ w-1\}=0. \tag{5.7}$$

Now χ appearing in (5.2) dominates $\|D^2u\|$, and so $\bar H$ is dominated by

$$H\triangleq \nabla h\cdot\nabla u-\|D^2u\|^2. \tag{5.8}$$

But let $x^0\in\mathscr{C}$ be given and choose $\varepsilon>0$ such that the closed disk $\overline{B_{2\varepsilon}(x_0)}$ is contained in $\mathscr{C}$. Choose a positive integer N such that

$$|\nabla u^{\varepsilon_n}(x)|<1 \quad \forall n\geqq N,\quad x\in\overline{B_{2\varepsilon}(x^0)}.$$

From (4.1i), (4.2), and (4.3), we see that

$$u^{\varepsilon_n}-\Delta u^{\varepsilon_n}=h \quad \text{on } \overline{B_{2\varepsilon}(x^0)}.$$

According to Gilbarg and Trudinger (1983, Thm. 4.6, p. 6), for every $\alpha\in(0,1)$, $|u^{\varepsilon_n}|_{C^{2,\alpha}(B_\varepsilon(x^0))}$ is bounded uniformly in $n\geqq N$. Thus, on $B_\varepsilon(x^0)$, $D^2u^{\varepsilon_n}$ is continuous and converges uniformly to D^2u, $\chi=\|D^2u\|^2$, and $\bar H=H$. We conclude that (5.7) remains valid if $\bar H$ is replaced by H, i.e.,

$$\max\{w-\tfrac{1}{2}\Delta w-H,\ w-1\}=0. \tag{5.9}$$

6. D^2u inside $\bar{\mathscr{C}}$. Inside the set $\mathscr{C}$ defined by (4.9), u satisfies the elliptic equation $u-\Delta u=h$, and is therefore smooth (at least $C^{4,\alpha}$ for all $\alpha\in(0,1)$ because h is $C^{2,1}$). In this section, we describe the behavior of D^2u as $\partial\mathscr{C}$ is approached from inside $\mathscr{C}$.

LEMMA 6.1. *Let* $z\in\partial\mathscr{C}$ *be given. As* $x\in\mathscr{C}$ *approaches* z, $D^2u(x)$ *approaches the matrix*

$$A(z)\triangleq(u(z)-h(z))\begin{bmatrix} u_2^2(z) & -u_1(z)u_2(z)\\ -u_1(z)u_2(z) & u_1^2(z)\end{bmatrix},$$

where u_i *denotes the ith partial derivative of* u.

Proof. Because $w=|\nabla u|^2=1$ on $\partial\mathscr{C}$, $A(z)$ can be characterized as the unique 2×2 positive semidefinite matrix with eigenvalues zero and $u(z)-h(z)$, and with $\nabla u(z)$ an eigenvector corresponding to the eigenvalue zero. Let v be a unit vector orthogonal to the unit vector $\nabla u(z)$. It suffices to show that

$$\lim_{\substack{x\to z\\ x\in\mathscr{C}}} D^2u(x)\nabla u(z)=0 \tag{6.1}$$

$$\lim_{\substack{x\to z\\ x\in\mathscr{C}}} D^2u(x)v=(u(z)-h(z))v. \tag{6.2}$$

Because $w = |\nabla u|^2$ attains its maximum value of 1 at z, and ∇w is continuous (Corollary 5.3), we have

$$0 = \nabla w(z) = \lim_{\substack{x \to z \\ x \in \mathscr{C}}} \nabla w(x) = \lim_{\substack{x \to z \\ x \in \mathscr{C}}} D^2 u(x) \nabla u(x).$$

Since ∇u is continuous and $D^2 u \in L^\infty_{loc}$, (6.1) follows.

Let $0 = \lambda_1(x) \leqq \lambda_2(x)$ denote the eigenvalues of $D^2 u(x)$. Then $u(x) - h(x) = \Delta u(x) = \lambda_1(x) + \lambda_2(x)$ for all $x \in \mathscr{C}$, and (6.1) shows that $\lim_{x \to z, x \in \mathscr{C}} \lambda_1(x) = 0$. Consequently,

$$\lim_{\substack{x \to z \\ x \in \mathscr{C}}} \lambda_2(x) = u(z) - h(z), \tag{6.3}$$

which is thus nonnegative. If $u(z) - h(z) = 0$, then $D^2 u(x)$ approaches the zero matrix and (6.2) holds. If $u(z) - h(z) > 0$, then (6.1) implies that any unit eigenvector corresponding to $\lambda_1(x)$ must, as $x \in \mathscr{C}$ approaches z, approach colinearity with $\nabla u(z)$. Hence, any unit eigenvector corresponding to $\lambda_2(x)$ approaches colinearity with ν, and (6.2) follows from (6.3). □

Remark 6.2. The characterization of $A(z)$ used in the proof of Lemma 6.1 makes critical use of the fact that our problem is posed in two dimensions. The two-dimensional nature of the problem also plays a fundamental role in Lemma 8.1, and together these lemmas provide the basis for § 10, where the existence of a continuous version of $D^2 u$ on $\mathbb{R}^2$ is established.

THEOREM 6.3. *For every $\alpha \in (0,1)$, $u \in C^{2,\alpha}(\bar{\mathscr{C}})$, i.e., $D^2 u$ restricted to $\mathscr{C}$ has an α-Hölder continuous extension to $\bar{\mathscr{C}}$.*

Proof. Because $|\nabla u| = 1$ on $\partial \mathscr{C}$, we can choose an open set $G \subset \mathscr{C}$ such that $|\nabla u|$ is bounded away from zero on $\mathscr{C} \setminus G$. Elliptic regularity implies the Hölder continuity of $D^2 u$ on $\bar{G}$, so it suffices to prove uniform Hölder continuity of $D^2 u$ on $\mathscr{C} \setminus G$.

Let a unit vector ν be given, and define on $\mathscr{C} \setminus G$, $\eta \triangleq \nabla u / |\nabla u|$, $z \triangleq \nu - (\nu \cdot \eta)\eta$,

$$\gamma \triangleq \begin{cases} \dfrac{z}{|z|} & \text{if } z \neq 0, \\ \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \eta & \text{if } z = 0. \end{cases}$$

Observe that $\eta \cdot \gamma = 0$ and $|\eta| = |\gamma| = 1$. Therefore,

$$\Delta u = (D^2 u)\eta \cdot \eta + (D^2 u)\gamma \cdot \gamma \quad \text{on } \mathscr{C} \setminus G.$$

Direct calculation shows that on $\mathscr{C} \setminus G$,

$$\begin{aligned} (D^2 u)\nu \cdot \nu &= (D^2 u) z \cdot z + 2(\nu \cdot \eta)(D^2 u)\eta \cdot z + (\nu \cdot \eta)^2 (D^2 u)\eta \cdot \eta \\ &= |z|^2 (\Delta u - (D^2 u)\eta \cdot \eta) + 2(\nu \cdot \eta)(D^2 u)\eta \cdot (\nu - (\nu \cdot \eta)\eta) \\ &\quad + (\nu \cdot \eta)^2 (D^2 u)\eta \cdot \eta. \end{aligned}$$

Since $\Delta u = u - h$ and $2(D^2 u)\eta = (\nabla w / |\nabla u|)$ on $\mathscr{C} \setminus G$, we have

$$\begin{aligned} (D^2 u)\nu \cdot \nu &= |\nu - \frac{(\nu \cdot \nabla u)}{|\nabla u|^2} \nabla u|^2 \left(u - h - \frac{1}{2} \frac{(\nabla w \cdot \nabla u)}{|\nabla u|^2} \right) \\ &\quad + \frac{(\nu \cdot \nabla u)}{|\nabla u|^2} \nabla w \cdot \nu - \frac{(\nu \cdot \nabla u)^2}{2|\nabla u|^4} \nabla w \cdot \nabla u \quad \text{on } \mathscr{C} \setminus G. \end{aligned} \tag{6.4}$$

All the terms appearing on the right-hand side of (6.4) are uniformly Hölder continuous in $\mathscr{C}\backslash G$ (recall Corollary 5.3). □

7. The gradient flow. Recalling Remark 4.6, we let $y^0 \in \mathscr{C}$ denote the unique minimizer of u. Using the strict convexity of u in $\mathscr{C}$ (Lemma 4.4), we choose $\delta > 0$, $\mu > 0$ such that

$$B_{2\delta}(y^0) \subset \mathscr{C}, \tag{7.1}$$

$$D^2u(x)y \cdot y \geqq \mu |y|^2 \quad \forall x \in B_{2\delta}(y^0), \tag{7.2}$$

$$\mu \leqq |\nabla u(x)|^2 \leqq \tfrac{1}{2} \quad \forall x \in \partial B_\delta(y^0), \tag{7.3}$$

$$\nabla u(y^0 + \delta\theta) \cdot \theta \geqq \mu \quad \forall \theta \in S_1, \tag{7.4}$$

where $S_1 \triangleq \partial B_1(0)$ is the set of unit vectors in $\mathbb{R}^2$. For $\theta \in S_1$, we define the *gradient flow* $\psi(t, \theta)$ to be the unique solution to the differential equation

$$\frac{d}{dt}\psi(t, \theta) = \nabla u(\psi(t, \theta)), \qquad t \geqq 0, \tag{7.5}$$

with the initial condition

$$\psi(0, \theta) = y^0 + \delta\theta. \tag{7.6}$$

We will find it convenient to use ψ to change coordinates in $\mathbb{R}^2$. The following theorem justifies this.

THEOREM 7.1. *The map ψ is a homeomorphism from $[0, \infty] \times S_1$ onto $\mathbb{R}^2 \backslash B_\delta(y^0)$.*

Proof. Let us for the moment fix $\theta \in S_1$ and define $n(t) \triangleq \psi(t, \theta) - y^0$ for all $t \geqq 0$. Because $|\nabla u| \leqq 1$, we have $|n(t)| \leqq t + \delta$, and $y^0 + (\delta \wedge t/t)n(t) \in B_{2\delta}(y^0)$ for all $t > 0$. We conclude from the convexity of u on $\mathbb{R}^2$ and from (7.2) that for $t > 0$:

$$\begin{aligned}
\frac{d}{dt}|n(t)|^2 &= 2\nabla u(y^0 + n(t)) \cdot n(t) \\
&= 2\left[\nabla u(y^0 + n(t)) - \nabla u\left(y^0 + \frac{\delta \wedge t}{t} n(t)\right)\right] \cdot n(t) \\
&\quad + 2\left[u\left(y^0 + \frac{\delta \wedge t}{t} n(t)\right) - \nabla u(y^0)\right] \cdot n(t) \qquad (7.7) \\
&\geqq 2\int_0^{\delta \wedge t/t} D^2u(y^0 + \tau n(t))n(t) \cdot n(t)\, d\tau \\
&\geqq 2\mu\left(1 \wedge \frac{\delta}{t}\right)|n(t)|^2.
\end{aligned}$$

Since $|n(0)|^2 = \delta^2$, we can integrate (7.7) to obtain the inequality

$$|\psi(t, \theta) - y^0|^2 \geqq \delta^2\left(1 \vee \frac{t}{\delta}\right)^{2\mu\delta} e^{2\mu(t \wedge \delta)} \quad \forall t \geqq 0, \quad \theta \in S_1. \tag{7.8}$$

One consequence of (7.8) is that

$$|\psi(s, \theta) - \psi(0, \varphi)| > 0 \quad \forall s > 0, \quad \theta \in S_1, \quad \varphi \in S_1. \tag{7.9}$$

Now let $s,\ t \in [0,\infty)$ and $\theta,\ \varphi \in S_1$ be given. Again using the convexity of u, we may write

$$
\begin{aligned}
|\psi(t+s,\theta)-\psi(t,\varphi)|^2 &= |\psi(s,\theta)-\psi(0,\varphi)|^2 \\
&\quad +2\int_0^t [\nabla u(\psi(\tau+s,\theta))-\nabla u(\psi(\tau,\varphi))]\cdot[\psi(\tau+s,\theta)-\psi(\tau,\varphi)]\,d\tau \\
&\geqq |\psi(s,\theta)-\psi(0,\varphi)|^2.
\end{aligned} \tag{7.10}
$$

If θ, φ are in S_1 and t_1, t_2 are in $[0,\infty)$ and $t_1 \neq t_2$, then (7.9), (7.10) imply that $\psi(t_1,\theta) \neq \psi(t_1,\varphi)$. If $t_1 = t_2$ but $\theta \neq \varphi$, then the uniqueness of solutions to (7.5) implies that $\psi(t_1,\theta) \neq \psi(t_2,\varphi)$. This concludes the proof that ψ is injective.

It is clear from its definition that ψ is continuous. Define

$$D \triangleq \psi([0,\infty)\times S_1) \subset \mathbb{R}^2 \backslash B_\delta(y^0)$$

to be the range of ψ. Let $x \in D$ and $\varepsilon > 0$ be given. It follows from (7.8) that there exists $T > 0$ such that

$$D \cap B_\varepsilon(x) \subset \psi([0,T]\times S_1).$$

But an injective, continuous map on a compact set has a continuous inverse, so ψ^{-1} is continuous at x.

It remains to show that $D = \mathbb{R}^2 \backslash B_\delta(y^0)$. There is a function $\hat{\psi}:[0,\infty)\times\mathbb{R}\to\mathbb{R}^2$ such that

$$\hat{\psi}(t,\beta) = \psi(t,(\cos\beta,\sin\beta)) \quad \forall (t,\beta)\in[0,\infty)\times\mathbb{R},$$

and $\hat{\psi}$ is continuous and locally injective. It follows from Deimling (1985, Thm. 4.3, p. 23) that

$$D \cap (\mathbb{R}^2 \backslash \overline{B_\delta(y^0)}) = \hat{\psi}((0,\infty)\times\mathbb{R})$$

is open. On the other hand, if $\{x^n\}_{n=1}^\infty \subset D$ is a sequence with limit $x^0 \in \mathbb{R}^2$, then (7.8) shows that $\{\psi^{-1}(x^n)\}_{n=1}^\infty$ is bounded and thus has an accumulation point $(t^0,\theta^0) \in [0,\infty)\times S_1$. The continuity of ψ implies that $x^0 = \psi(t^0,\theta^0)$, so D is closed. It follows that $D = \mathbb{R}^2 \backslash B_\delta(y^0)$. □

COROLLARY 7.2. *For $\theta \in S_1$ and $\gamma \in [\frac{1}{2},1]$, define*

$$T_\gamma(\theta) \triangleq \inf\{t \geqq 0;\ |\nabla u(\psi(t,\theta))|^2 \geqq \gamma\}. \tag{7.11}$$

Then

$$\sup_{\substack{1/2\leqq\gamma\leqq 1\\ \theta\in S_1}} T_\gamma(\theta) \leqq \sup_{\theta\in S_1} T_1(\theta) < \infty.$$

Proof. According to Lemma 4.4, $\mathscr{C}$ is bounded. We can use (7.8) to choose $t^* \in (0,\infty)$ such that

$$\mathscr{C} \subset \psi([0,t^*]\times S^1). \qquad \square$$

THEOREM 7.3. *The homeomorphism ψ is Lipschitz continuous on compact subsets of $[0,\infty)\times S_1$, and ψ^{-1} is Lipschitz continuous on all of $\mathbb{R}^2 \backslash B_\delta(y^0)$.*

Proof. It follows immediately from (7.5) that $|(d/dt)\psi(t,\theta)| \leqq 1$ for all $(t,\theta) \in [0,\infty)\times S_1$. Now let $T>0$ be given and use Theorem 4.3 to choose a Lipschitz constant C for ∇u on $\psi([0,T]\times S_1)$. For $\theta, \varphi \in S_1$ and $t \in [0,T]$, we have

$$\begin{aligned}|\psi(t,\theta)-\psi(t,\varphi)| &\leqq |\psi(0,\theta)-\psi(0,\psi)| \\ &\quad + \int_0^t |\nabla u(\psi(\tau,\theta))-\nabla u(\psi(\tau,\varphi))|\,d\tau \\ &\leqq \delta|\theta-\varphi| + C\int_0^t |\psi(\tau,\theta)-\psi(\tau,\varphi)|\,d\tau.\end{aligned}$$

Gronwall's inequality gives

$$|\psi(t,\theta)-\psi(t,\varphi)| \leqq \delta\, e^{CT}|\theta-\varphi|,$$

and the local Lipschitz continuity of ψ is proved.

To prove the global Lipschitz continuity of ψ^{-1}, we let $x^1, x^2 \in \mathbb{R}^2\backslash B_\delta(y^0)$ be given and define $(t_1,\theta_1)=\psi^{-1}(x^1)$, $(t_2,\theta_2)=\psi^{-1}(x^2)$. Assume without loss of generality that $|x^1-x^2| \leqq 1$ and that $t_1 \geqq t_2$. Set $s=t_1-t_2$. According to (7.10) and (7.8),

$$\begin{aligned}|x^1-x^2| &\geqq |\psi(s,\theta_1)-\psi(0,\theta_2)| \\ &\geqq |\psi(s,\theta_1)-y^0| - |y^0-\psi(0,\theta_2)| \\ &\geqq \delta\left(1\vee\frac{s}{\delta}\right)^{\mu\delta} e^{\mu(s\wedge\delta)} - \delta \\ &\geqq \delta\mu\left(1\vee\frac{s}{\delta}\right)^{\mu\delta}(s\wedge\delta).\end{aligned} \tag{7.12}$$

If $0 \leqq s \leqq \delta$, then (7.12) yields

$$|t_1-t_2| \leqq \frac{1}{\delta\mu}|x^1-x^2|. \tag{7.13}$$

If $s \geqq \delta$ and $\mu\delta \geqq 1$, (7.12) again yields (7.13). Finally, if $s \geqq \delta$ and $0<\mu\delta<1$, (7.12) yields $|x^1-x^2| \geqq \mu\delta^{1-\mu\delta}s^{\mu\delta}$, so

$$|t_1-t_1| \leqq (\mu\delta^{1-\mu\delta})^{1/\mu\delta}|x^1-x^2|^{1/\mu\delta} \leqq (\mu\delta^{1-\mu\delta})^{1/\mu\delta}|x^1-x^2|. \tag{7.14}$$

Relations (7.14) and (7.15) imply the global Lipschitz continuity of the first component of ψ^{-1}, i.e., there exists a constant $L>0$ such that

$$|t_1-t_2| \leqq L|\psi(t_1,\theta_1)-\psi(t_2,\theta_2)| \quad \forall (t_1,\theta_1),(t_2,\theta_2) \in [0,\infty)\times S_1. \tag{7.15}$$

Now let $x^1, x^2 \in \mathbb{R}^2\backslash B_\delta(y^0)$ be given, and define (t_1,θ_1), (t_2,θ_2), and $s=t_1-t_2 \geqq 0$ as before. From (7.10), (7.5), and (7.6), we have

$$\begin{aligned}|x^1-x^2| &\geqq |\psi(s,\theta_1)-\psi(0,\theta_2)| \\ &\geqq -|\psi(s,\theta_1)-\psi(0,\theta_1)| + |\psi(0,\theta_1)-\psi(0,\theta_2)| \\ &\geqq -s+\delta|\theta_1-\theta_2|.\end{aligned}$$

Relation (7.15) gives us

$$|\theta_1-\theta_2| \leqq \frac{1}{\delta}|t_1-t_2| + \frac{1}{\delta}|x^1-x^2| \leqq \frac{1}{\delta}(1+L)|x^1-x^2|. \qquad \square$$

Remark 7.4. In much of what follows, we will use the coordinates $(t, \theta) \in [0, \infty) \times S_1$ rather than the coordinates $x \in \mathbb{R}^2 \setminus B_\delta(y^0)$. We may identify S_1 with the unit circle, and let $[0, \infty) \times S_1$ have the product of Lebesgue measure and arc length measure. An important consequence of Theorem 7.3 is that ψ maps measure zero subsets of $[0, \infty) \times S_1$ onto Lebesgue measure zero subsets of $\mathbb{R}^2 \setminus B_\delta(y^0)$. Likewise, ψ^{-1} preserves measure zero sets.

8. $W^{2,\infty}$ regularity for the obstacle problem. The purpose of this section is to show that the function $w = |\nabla u|^2$ is in $W^{2,\infty}_{loc}$. This improves the regularity result of Theorem 5.2.

LEMMA 8.1. *We have*

$$(D^2u)\nabla u = 0, \quad \|D^2u\| = \Delta u \quad \text{a.e. on } \mathbb{R}^2 \setminus \mathscr{C}. \tag{8.1}$$

Proof. By the definition of $\mathscr{C}$, w attains its maximum value of 1 at every point in $\mathbb{R}^2 \setminus \mathscr{C}$, so $\nabla w = 0$ everywhere on $\mathbb{R}^2 \setminus \mathscr{C}$. But $\nabla w = 2(D^2u)\nabla u$ almost everywhere on $\mathbb{R}^2$, and the first part of (8.1) follows. Since D^2u is singular almost everywhere on $\mathbb{R}^2 \setminus \mathscr{C}$, the second part of (8.1) also holds. □

Remark 8.2. Because D^2u is positive definite on $\mathscr{C}$ and positive semidefinite almost everywhere on $\mathbb{R}^2$, and since (recalling Remark 7.4)

$$\frac{d}{dt} w(\psi(t, \theta)) = 2D^2u(\psi(t, \theta))\nabla u(\psi(t, \theta)) \cdot \nabla u(\psi(t, \theta)) \tag{8.2}$$

$$\text{a.e. } (t, \theta) \in [0, \infty) \times S^1,$$

the function $t \mapsto w(\psi(t, \theta))$ is nondecreasing for almost every $\theta \in S^1$. In particular, with $T_1(\theta)$ defined by (7.11), we have

$$w(\psi(t, \theta)) \equiv 1 \quad \forall t \geqq T_1(\theta), \quad \text{a.e. } \theta \in S^1. \tag{8.3}$$

THEOREM 8.3. *The function $w = |\nabla u|^2$ is in $W^{2,\infty}$.*

Proof. Recall that w satisfies (5.9), where for all $\alpha \in (0, 1)$, $H \triangleq \nabla h \cdot \nabla u - \|D^2u\|^2$ is of class $C^{0,\alpha}$ inside $\mathscr{C}$, and H is defined up to almost everywhere equivalence on $\mathbb{R}^2 \setminus \mathscr{C}$. We define

$$\hat{H}(x) \triangleq \begin{cases} \nabla h(x) \cdot \nabla u(x) - \|D^2u(x)\|^2 & \forall x \in \mathscr{C} \\ \nabla h(x) \cdot \nabla u(x) - [(u(x) - h(x))^+]^2 & \text{if } x \in \mathbb{R}^2 \setminus \mathscr{C}. \end{cases} \tag{8.4}$$

Now $u - h = \Delta u \geqq 0$ on $\mathscr{C}$, so $u - h \geqq 0$ on $\partial\mathscr{C}$. Theorem 6.3 and Lemma 6.1 then show that $\hat{H}$ is locally Hölder continuous with exponent α for any $\alpha \in (0, 1)$. Because of (3.1) and Lemma 8.1,

$$u - h \leqq \Delta u = \|D^2u\| \quad \text{a.e. on } \mathbb{R}^2 \setminus \mathscr{C}.$$

But $\Delta u \geqq 0$ almost everywhere $\mathbb{R}^2$, so

$$[(u - h)^+]^2 \leqq \|D^2u\|^2 \quad \text{a.e. on } \mathbb{R}^2 \setminus \mathscr{C}.$$

Therefore $\hat{H} \geqq H$ amost everywhere $\mathbb{R}^2 \setminus \mathscr{C}$, and $\hat{H} = H$ on $\mathscr{C}$, so (5.9) yields

$$\max\{w - \tfrac{1}{2}\Delta w - \hat{H}, w - 1\} = 0. \tag{8.5}$$

With the aid of (8.5) and the Hölder continuity of $\hat{H}$, we can obtain the $W^{2,\infty}$ regularity of w from the theory of variational inequalities. More precisely, choose r so that $\mathscr{C} \subset B_r(0)$ and observe that the Dirichlet problem

$$\varphi - \tfrac{1}{2}\Delta\varphi = \hat{H} \quad \text{on } B_r(0), \quad \varphi = 0 \quad \text{on } \partial B_r(0).$$

has a solution φ which is in $C^{2,\alpha}(\overline{B_r(0)})$ for any $\alpha \in (0,1)$ (Ladyzhenskaya and Ural'tseva (1968, Thm. 3.1.3, p. 115)). Set $\bar{w} \triangleq w - \varphi$, so that $\bar{w} \in W^{2,p}(\overline{B_r(0)})$ for any $p \in (1,\infty)$ and

$$\max\{\bar{w} - \tfrac{1}{2}\Delta\bar{w}, \bar{w} - 1 + \varphi\} = 0 \quad \text{in } B_r(0), \tag{8.6}$$

$$\bar{w} = 1 \quad \text{on } \partial B_r(0). \tag{8.7}$$

Define

$$L_r \triangleq \{v \in W^{1,2}(B_r(0)); -\varphi \leqq v \leqq 1 - \varphi \text{ on } B_r(0) \text{ and } v - 1 \in W_0^{1,2}(B_r)\},$$

and note from (8.6), (8.7) that $\bar{w} \in L_r$ and

$$\frac{1}{2}\int_{B_r(0)} \nabla\bar{w}\cdot(\nabla v - \nabla\bar{w}) \geqq -\int_{B_r(0)} \bar{w}(v - \bar{w}) \quad \forall v \in L_r.$$

It follows from Chipot (1984, Thm. 3.25, p. 49), that $\bar{w} \in W^{2,\infty}(B_r(0))$, so also $w \in W^{2,\infty}(B_r(0))$. On $\mathbb{R}^2 \setminus B_r(0)$, $w \equiv 1$. □

COROLLARY 8.4. *We have* $D^2u \in W^{1,\infty}(\bar{\mathscr{C}})$.

Proof. Use the $W^{1,\infty}$ regularity of ∇w in (6.4). □

9. Lipschitz continuity of T_γ. Recall the mappings $T_\gamma : S_1 \mapsto [0,\infty)$ defined by (7.11) for each $\gamma \in [\frac{1}{2}, 1]$. The continuity of $\nabla u \circ \psi$ implies the lower semicontinuity of each T_γ. In this section we prove that for each $\gamma \in [\frac{1}{2}, 1]$, T_γ is, in fact, Lipschitz continuous.

LEMMA 9.1. *We have*

$$K \triangleq \sup_{\nu \in S_1, x \in \mathscr{C}} \frac{|\nabla w(x)|}{D^2u(x)\nu\cdot\nu} < \infty. \tag{9.1}$$

Proof. Let ν, $\eta \in S_1$ be given and set $f \triangleq (D^2u)\nu\cdot\nu$ and $g \triangleq \nabla w \cdot \eta$. Then in $\mathscr{C}$,

$$f - \Delta f = (D^2h)\nu\cdot\nu \geqq c_0, \qquad g - \Delta g = 2\nabla H\cdot\eta - g,$$

where $c_0 > 0$ is the constant in (2.8), and H, defined by (5.8), is in $W^{1,\infty}(\bar{\mathscr{C}})$ because of Corollary 8.4. Furthermore, $g = 0 \leqq f$ on $\partial\mathscr{C}$. Therefore the maximum principle implies that $g - Kf \leqq 0$ in $\mathscr{C}$, where

$$K \triangleq \frac{1}{c_0}(2\|\nabla H\|_{L^\infty(\bar{\mathscr{C}})} + \|\nabla w\|_{L^\infty(\bar{\mathscr{C}})}).$$

In other words, $\nabla w\cdot\eta \leqq K(D^2u)\nu\cdot\nu$. □

THEOREM 9.2. *For each $\gamma \in [\frac{1}{2}, 1]$, the mapping $T_\gamma : S_1 \mapsto [0,\infty)$ is Lipschitz continuous with a Lipschitz constant which is independent of γ.*

Proof. For each $\gamma \in [\frac{1}{2}, 1]$, define

$$\mathscr{C}_\gamma \triangleq \{\psi(t,\theta); 0 \leqq t < T_\gamma(\theta)\} \cup B_\delta(y^0)$$

(with ψ, δ, and y^0 as in (7.1)–(7.6)). Each $\mathscr{C}_\gamma$ is open, $w < \gamma$ on $\mathscr{C}_\gamma$ and $w = \gamma$ on $\partial\mathscr{C}_\gamma$. For $\gamma \in [\frac{1}{2}, 1)$, we also have $\mathscr{C}_\gamma \subset \mathscr{C}$. Because of (4.10), ∇w does not vanish on $\mathscr{C}$, so for fixed $\gamma \in [\frac{1}{2}, 1)$ and $z \in \partial\mathscr{C}_\gamma$, the outward normal to $\mathscr{C}_\gamma$ exists and is

$$n(z) \triangleq \frac{\nabla w(z)}{|\nabla w(z)|} = \frac{2D^2u(z)\nabla u(z)}{|\nabla w(z)|}.$$

In fact D^2w is continuous in $\mathscr{C}$ and bounded in $\mathbb{R}^2$ (Theorem 8.3), so for every $\gamma \in [\frac{1}{2}, 1)$, $\partial\mathscr{C}_\gamma$ has bounded curvature, i.e., there are constants $\varepsilon > 0$, $K_\gamma > 0$ such that for every $z \in \partial\mathscr{C}_\gamma$, and for every $x \in B_\varepsilon(z)$:

$$(x - z)\cdot n(z) \geqq K_\gamma |x - z|^2 \Rightarrow x \in \mathbb{R}^2 \setminus \mathscr{C}_\gamma. \tag{9.2}$$

We may use the local boundedness of $(d^2/dt^2)\psi(t,\theta)=\frac{1}{2}\nabla w(\psi(t,\theta))$ and the Lipschitz continuity of ψ to choose a constant $K_2>0$ such that for every $\gamma\in[\frac{1}{2},1)$, every $\beta\in[0,1]$, and every $\theta,\varphi\in S_1$:

$$|\psi(T_\gamma(\theta)+\beta,\theta)-\psi(T_\gamma(\theta),\theta)-\beta\nabla u(\psi(T_\gamma(\theta),\theta))|\leqq K_2\beta^2, \tag{9.3}$$

$$|\psi(T_\gamma(\theta)+\beta,\theta)-\psi(T_\gamma(\theta)+\beta,\varphi)|\leqq K_2|\theta-\varphi|. \tag{9.4}$$

With K as in (9.1), choose $L>\max\{\frac{1}{2}KK_2,1\}$. Let $\theta,\varphi\in S_1$ be given with $|\theta-\varphi|\leqq 1/L$, and set

$$\beta=L|\theta-\varphi|,\quad z=\psi(T_\gamma(\theta),\theta),\quad x=\psi(T_\gamma(\theta)+\beta,\varphi).$$

Then (9.3), (9.4) imply the existence of vectors $\nu,\eta\in B_1(0)$ such that

$$x=z+\beta\nabla u(z)+K_2\beta^2\nu+K_2|\theta-\varphi|\eta.$$

We calculate

$$\begin{aligned}(x-z)\cdot n(z)&=\frac{2\beta D^2u(z)\nabla u(z)\cdot\nabla u(z)}{|\nabla w(z)|}+K_2\beta^2 n(z)\cdot\nu+K_2|\theta-\varphi|n(z)\cdot\eta\\&\geqq\frac{2\beta}{K}-K_2\beta^2-K_2|\theta-\varphi|\\&=\left(\frac{2L}{K}-K_2\right)|\theta-\varphi|-K_2L^2|\theta-\varphi|^2,\end{aligned}$$

and

$$\begin{aligned}K_\gamma|x-z|^2&=K_\gamma|\beta\nabla u(z)+K_2\beta^2\nu+K_2|\theta-\varphi|\eta|^2\\&\leqq 9K_\gamma(L^2+K_2^2L^4+K_2^2)|\theta-\varphi|^2.\end{aligned}$$

It is clear that for $|\theta-\varphi|$ sufficiently small, $x\in B_\varepsilon(z)$ and

$$(x-z)\cdot n(z)\geqq K_\gamma|x-z|^2,$$

from which we conclude (see (9.2)) that $x\in\mathbb{R}^2\setminus\mathscr{C}_\gamma$, i.e.,

$$T_\gamma(\varphi)\leqq T_\gamma(\theta)+\beta=T_\gamma(\theta)+L|\theta-\varphi|.$$

Interchanging the roles of θ and φ, we obtain

$$|T_\gamma(\theta)-T_\gamma(\varphi)|\leqq L|\theta-\varphi|$$

for all $\theta,\varphi\in S_1$ such that $|\theta-\varphi|$ is sufficiently small.

For each $\theta\in S_1$, the mapping $t\mapsto w(\psi(t,\theta))$ is strictly increasing on $[0,T_1(\theta)]$ (see (8.2) and (4.10)). Therefore, the mapping $\gamma\mapsto T_\gamma(\theta)$ is continuous on $[\frac{1}{2},1]$. The Lipschitz continuity of T_1 follows from the uniform Lipschitz continuity of T_γ for $\gamma\in[\frac{1}{2},1)$. □

COROLLARY 9.3. *With ψ, δ, and y^0 as in* (7.1)-(7.6), *we have*

$$\mathscr{C}=\{\psi(t,\theta);\ \theta\in S^1,\ t\in[0,T_1(\theta))\}\cup B_\delta(y^0). \tag{9.5}$$

Proof. Define $\tilde{\mathscr{C}}$ to be the set on the right-hand side of (9.5). It is clear that $\tilde{\mathscr{C}}\subset\mathscr{C}$, and because of (8.3) and Remark 7.4, the Lebesgue measure of $\mathscr{C}\setminus\tilde{\mathscr{C}}$ is zero. Let $x\in\mathscr{C}\setminus\tilde{\mathscr{C}}$ be given, and define $(t,\theta)\triangleq\psi^{-1}(x)$. Then $t\geqq T_1(\theta)$, but because $w(T_1(\theta),\theta)=1$, we must in fact have $t>T_1(\theta)$. The continuity of T_1 and w allows us to chose an open neighborhood of (t,θ) contained in $\mathscr{C}\setminus\tilde{\mathscr{C}}$, and this contradicts the Lebesgue negligibility of $\mathscr{C}\setminus\tilde{\mathscr{C}}$. □

10. D^2u outside $\mathscr{C}$. We saw in Lemma 8.1 that D^2u is singular almost everywhere in $\mathbb{R}^2\setminus\mathscr{C}$. Indeed

$$u_{11}u_1+u_{12}u_2=0, \quad u_{12}u_1+u_{22}u_2=0 \quad \text{a.e. on } \mathbb{R}^2\setminus\mathscr{C}, \tag{10.1}$$

and because $u_1^2+u_2^2=1$ on $\mathbb{R}^2\setminus\mathscr{C}$, we have

$$D^2u=\Delta u\begin{bmatrix} u_2^2 & -u_1u_2 \\ -u_1u_2 & u_1^2 \end{bmatrix} \quad \text{a.e. on } \mathbb{R}^2\setminus\mathscr{C}. \tag{10.2}$$

Because u has continuous first partial derivatives on $\mathbb{R}^2$, the proof of continuity of D^2u on $\mathbb{R}^2\setminus\mathscr{C}$ reduces to a search for a continuous version of Δu on this set. In order for D^2u to be continuous across $\partial\mathscr{C}$, we must also have $\Delta u=u-h$ on $\partial\mathscr{C}$ (see Lemma 6.1).

We shall construct the desired continuous version of Δu in the (t,θ) variables. Indeed, if we set

$$\lambda(t,\theta)=\Delta u(\psi(t,\theta)) \quad \forall\theta\in S^1, \quad t\geqq T_1(\theta),$$

then a formal calculation relying on (10.2) and the constancy of w on $\mathbb{R}^2\setminus\mathscr{C}$ leads to

$$\begin{aligned}\frac{d}{dt}\lambda(t,\theta)&=\frac{1}{2}\Delta w(\psi(t,\theta))-\|D^2u(\psi(t,\theta))\|^2\\ &=-\lambda^2(t,\theta) \quad \forall\theta\in S^1, \quad t\geqq T_1(\theta).\end{aligned} \tag{10.3}$$

Integrating this equation and invoking the condition $\Delta u=u-h$ on $\partial\mathscr{C}$, we obtain

$$\lambda(t,\theta)=\frac{u(\psi(T_1(\theta),\theta))-h(\psi(T_1(\theta),\theta))}{1+(t-T_1(\theta))[u(\psi(T_1(\theta),\theta))-h(\psi(T_1(\theta),\theta))]} \tag{10.4}$$

$$\forall\theta\in S_1, \quad t\geqq T_1(\theta).$$

The task before us is to show that with λ defined by (10.4), the function $\lambda\circ\psi^{-1}$ is a version of Δu on $\mathbb{R}^2\setminus\mathscr{C}$. This is essentially a justification of the formal differentiation in (10.3), which involved third-order derivatives of u.

Let $\rho:\mathbb{R}^2\to[0,\infty)$ be a C^∞ function with support in $B_1(0)$ and satisfying $\int_{\mathbb{R}^2}\rho=1$. For $n=1,2,\cdots$, we define mollifications of u by

$$u^{(n)}(x)\triangleq\int_{\mathbb{R}^2}u\left(x-\frac{1}{n}\xi\right)\rho(\xi)\,d\xi=n^2\int_{\mathbb{R}^2}u(\xi)\rho(n(x-\xi))\,d\xi. \tag{10.5}$$

Then $\nabla u^{(n)}$ and $D^2u^{(n)}$ are locally bounded, uniformly in n, and $u^{(n)}\to u$, $\nabla u^{(n)}\to\nabla u$, and $D^2u^{(n)}\to D^2u$ in L^1_{loc}. By passing to subsequences if necessary, we assume that these convergences occur almost everywhere. We define for $(t,\theta)\in[0,\infty)\times S^1$:

$$l^{(n)}(t,\theta)\triangleq\Delta u^{(n)}(\psi(t,\theta)), \qquad n=1,2,\cdots, \tag{10.6}$$

$$l(t,\theta)\triangleq\Delta u(\psi(t,\theta)), \tag{10.7}$$

and observe that $l^{(n)}(t,\theta)\to l(t,\theta)$ for almost every $(t,\theta)\in[0,\infty)\times S^1$ (Remark 7.4).

LEMMA 10.1. *The functions*

$$\dot{l}^{(n)}(t,\theta)=\nabla\Delta u^{(n)}(\psi(t,\theta))\cdot\nabla u(\psi(t,\theta)) \tag{10.8}$$

are locally bounded, uniformly in n.

Proof. Observe first of all that

$$\begin{aligned}\dot{l}^{(n)}&=\nabla\Delta u^{(n)}\cdot\nabla u^{(n)}+\nabla\Delta u^{(n)}\cdot(\nabla u-\nabla u^{(n)})\\ &=\tfrac{1}{2}\Delta(|\nabla u^{(n)}|^2)-\|D^2u^{(n)}\|^2+\nabla\Delta u^{(n)}\cdot(\nabla u-\nabla u^{(n)}),\end{aligned}$$

where $\dot{l}^{(n)}$ is evaluated at (t, θ), and the right-hand side is evaluated at $\psi(t, \theta)$. It suffices to obtain uniform local bounds on $\Delta(|\nabla u^{(n)}|^2)$ and $\nabla \Delta u^{(n)} \cdot (\nabla u - \nabla u^{(n)})$.

Define for $i \in \{1, 2\}$ the functions

$$\begin{aligned} F_{ii}^{(n)}(x) &\triangleq \int_{\mathbb{R}^2} \left(\left| \nabla u\left(x - \frac{1}{n}\xi\right)\right|^2 \right)_{i,i} \rho(\xi)\, d\xi \\ &= n \int_{\mathbb{R}^2} \left(\left| \nabla u\left(x - \frac{1}{n}\xi\right)\right|^2 \right)_{i} \rho_i(\xi)\, d\xi \\ &= 2n \int_{\mathbb{R}^2} \nabla u_i\left(x - \frac{1}{n}\xi\right) \cdot \nabla u\left(x - \frac{1}{n}\xi\right) \rho_i(\xi)\, d\xi, \end{aligned}$$

$$n = 1, 2, \cdots,$$

and note that these functions are uniformly bounded in n (Theorem 8.3). Then

$$\begin{aligned} (|\nabla u^{(n)}(x)|^2)_{ii} &= 2n^6 \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \nabla u(\xi) \cdot \nabla u(\eta) [\rho_{ii}(n(x-\xi))\rho(n(x-\eta)) \\ &\qquad + \rho_i(n(x-\xi))\rho_i(n(x-\eta))]\, d\xi\, d\eta \\ &= 2n^2 \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \nabla u\left(x - \frac{1}{n}\xi\right) \cdot \nabla u\left(x - \frac{1}{n}\eta\right) [\rho_{ii}(\xi)\rho(\eta) \\ &\qquad + \rho_i(\xi)\rho_i(\eta)]\, d\xi\, d\eta \\ &= 2n \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \nabla u_i\left(x - \frac{1}{n}\xi\right) \cdot \nabla u\left(x - \frac{1}{n}\eta\right) \rho_i(\xi)\rho(\eta)\, d\xi\, d\eta \\ &\quad + 2 \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \nabla u_i\left(x - \frac{1}{n}\xi\right) \nabla u_i\left(x - \frac{1}{n}\eta\right) \rho(\xi)\rho(\eta)\, d\xi\, d\eta. \end{aligned}$$

The last term is locally bounded in x, uniformly in n. The next-to-last term is

$$F_{ii}^{(n)}(x) + 2n \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \nabla u_i\left(x - \frac{1}{n}\xi\right) \cdot \left[\nabla u\left(x - \frac{1}{n}\eta\right) - \nabla u\left(x - \frac{1}{n}\xi\right)\right] \rho_i(\xi)\rho(\eta)\, d\xi\, d\eta,$$

which is also locally bounded in x, uniformly in n, because for all $\xi, \eta \in B_1(0)$,

$$\left| \nabla u\left(x - \frac{1}{n}\eta\right) - \nabla u\left(x - \frac{1}{n}\xi\right)\right| \leqq \frac{2}{n} \sup_{B_1(x)} \|D^2 u\|.$$

This provides a uniform local bound on $\Delta(|\nabla u^{(n)}|^2)$.

On the other hand,

$$\begin{aligned} &\nabla \Delta u^{(n)}(x) \cdot (\nabla u(x) - \nabla u^{(n)}(x)) \\ &\quad = n^5 \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \Delta u(\xi) \\ &\qquad \times [\nabla u(x) - \nabla u(\eta)] \cdot \nabla \rho(n(x-\xi))\rho(n(x-\eta))\, d\xi\, d\eta \\ &\quad = n \int_{\mathbb{R}^2} \int_{\mathbb{R}^2} \Delta u\left(x - \frac{1}{n}\xi\right) \\ &\qquad \times \left[\nabla u(x) - \nabla u\left(x - \frac{1}{n}\eta\right)\right] \cdot \nabla \rho(\xi)\rho(\eta)\, d\xi\, d\eta, \end{aligned}$$

and the boundedness of this expression follows from the local Lipschitz continuity of ∇u. □

Because of Lemma 10.1, a subsequence of $\{\dot{l}^{(n)}\}_{n=1}^{\infty}$ converges in the L^{∞}_{loc}-weak* topology to a function $\zeta \in L^{\infty}_{loc}([0,\infty)\times S_1)$. We assume without loss of generality that the full sequence converges. For each nonnegative integer k, choose a number $t_k > k$ such that $\{l^{(n)}(t_k,\theta)\}_{n=1}^{\infty}$ converges for almost every $\theta \in S_1$, and define $\lambda_k(t_k,\theta)$ to be this limit. (Whereas $l(\cdot,\cdot)$ is defined up to almost everywhere equivalence on $[0,\infty)\times S_1$, the functions $\lambda_k(t_k,\cdot)$ are defined up to almost everywhere equivalence on S_1.) We insist furthermore that t_0 be chosen so that $\psi(t_0,\theta)\in\mathscr{C}$ for all $\theta\in S_1$. Then $\Delta u(\psi(t_0,\cdot))$ is defined pointwise on S_1 because Δu is continuous on $\mathscr{C}$, and so we may require that

$$\lambda_0(t_0,\theta) = \Delta u(\psi(t_0,\theta)) \quad \forall \theta \in S_1.$$

For each $k = 0, 1, \cdots$, define $\lambda_k : [0,\infty)\times S^1 \mapsto \mathbb{R}$ by

$$\lambda_k(t,\theta) \triangleq \lambda_k(t_k,\theta) + \int_{t_k}^{t} \zeta(s,\theta)\, ds,$$

so that any two versions $\hat{\lambda}_k$ and $\tilde{\lambda}_k$ of this function have the property that the set $\{\theta\in S_1 |$ there exists $t\in[0,\infty]$ with $\hat{\lambda}_k(t,\theta)\neq\tilde{\lambda}_k(t,\theta)\}$ has measure zero.

We now relate the functions λ_k, $k=0,1,\cdots$, to the function l of (10.7). Let φ be a continuous, real-valued function on $[0,\infty)\times S_1$, and define

$$\Phi(t,\theta) \triangleq \int_0^t \varphi(s,\theta)\, ds \quad \forall (t,\theta)\in[0,\infty)\times S_1.$$

For $k = 0, 1, \cdots$,

$$\begin{aligned}
&\int_{S_1}\int_0^{t_k} \lambda_k(s,\theta)\varphi(s,\theta)\, ds\, d\theta \\
&\quad = \int_{S_1}\left[\lambda_k(t_k,\theta)\Phi(t_k,\theta) - \int_0^{t_k}\zeta(s,\theta)\varphi(s,\theta)\, ds\right] d\theta \\
&\quad = \lim_{n\to\infty}\int_{S_1}\left[l^{(n)}(t_k,\theta)\Phi(t_k,\theta) - \int_0^{t_k}\dot{l}^{(n)}(s,\theta)\varphi(s,\theta)\, ds\right] d\theta \\
&\quad = \lim_{n\to\infty}\int_{S_1}\int_0^{t_k} l^{(n)}(s,\theta)\varphi(s,\theta)\, ds\, d\theta \\
&\quad = \int_{S_1}\int_0^{t_k} l(s,\theta)\varphi(s,\theta)\, ds\, d\theta.
\end{aligned}$$

It follows that $\lambda_k = l$ almost everywhere on $[0,t_k]\times S_1$. In particular, for any two nonnegative integers k and m, λ_k and λ_m agree almost everywhere on $[0, t_k\wedge t_m]\times S_1$, and hence almost everywhere on $[0,\infty)\times S_1$. In particular,

$$\lambda_0(t,\theta) = \Delta u(\psi(t,\theta)), \quad \text{a.e. } (t,\theta)\in[0,\infty)\times S^1, \tag{10.9}$$

and for almost every $\theta\in S_1$,

$$\lambda_0(t,\theta) = \Delta u(\psi(t_0,\theta)) + \int_{t_0}^{t}\zeta(s,\theta)\, ds \quad \forall t\in[0,\infty). \tag{10.10}$$

LEMMA 10.2. *Almost everywhere on the set*

$$\psi^{-1}(\mathbb{R}^2\setminus\mathscr{C}) = \{(t,\theta)\in[0,\infty)\times S_1;\ t \geqq T_1(\theta)\},$$

the function ζ appearing in (10.10) *is equal to* $-\lambda_0^2$.

Proof. From (10.8) we have

$$\begin{aligned}\dot{l}^{(n)}\circ\psi^{-1}+(l\circ\psi^{-1})(l^{(n)}\circ\psi^{-1})&=\nabla\Delta u^{(n)}\cdot\nabla u+\Delta u\,\Delta u^{(n)}\\&=(u_{12}^{(n)}u_2+u_{11}^{(n)}u_1)_1+(u_{12}^{(n)}u_1+u_{22}^{(n)}u_2)_2\\&\quad+u_{11}^{(n)}u_{22}+u_{11}u_{22}^{(n)}-2u_{12}^{(n)}u_{12}.\end{aligned}$$

Now $u_{11}^{(n)}u_{22}+u_{11}u_{22}^{(n)}-2u_{12}^{(n)}u_{12}$ is locally bounded, uniformly in n, and converges almost everywhere to $2\det D^2u$, which is zero on $\mathbb{R}^2\setminus\bar{\mathscr{C}}$. It follows from (10.1) that for any function $\varphi\in C_0^1(\mathbb{R}^2\setminus\bar{\mathscr{C}})$,

$$\begin{aligned}&\lim_{n\to\infty}\int_{\mathbb{R}^2\setminus\bar{\mathscr{C}}}[\dot{l}^{(n)}\circ\psi^{-1}+(l\circ\psi^{-1})(l^{(n)}\circ\psi^{-1})]\varphi\\&\quad=\lim_{n\to\infty}\int_{R^2\setminus\bar{\mathscr{C}}}[(u_{12}^{(n)}u_2+u_{11}^{(n)}u_1)_1+(u_{12}^{(n)}u_1+u_{22}^{(n)}u_2)_2]\varphi\\&\quad=-\lim_{n\to\infty}\int_{\mathbb{R}^2\setminus\bar{\mathscr{C}}}(u_{12}^{(n)}u_2+u_{11}^{(n)}u_1)\varphi_1+(u_{12}^{(n)}u_1+u_{22}^{(n)}u_2)\varphi_2\\&\quad=0.\end{aligned}$$

Because the functions $\dot{l}^{(n)}\circ\psi^{-1}+(l\circ\psi^{-1})(l^{(n)}\circ\psi^{-1})$ are locally bounded, uniformly in n, we can show that for every $\varphi\in L^1(\mathbb{R}^2\setminus\bar{\mathscr{C}})$,

$$\lim_{n\to\infty}\int_{\mathbb{R}^2\setminus\bar{\mathscr{C}}}[\dot{l}^{(n)}\circ\psi^{-1}+(l\circ\psi^{-1})(l^{(n)}\circ\psi^{-1})]\varphi=0. \tag{10.11}$$

Now let $\gamma\in L^1(\psi^{-1}(\mathbb{R}^2\setminus\bar{\mathscr{C}}))$ be given so that $(\gamma\circ\psi^{-1})|J\Psi^{-1}|\in L^1(\mathbb{R}^2\setminus\bar{\mathscr{C}})$, where $|J\psi^{-1}|$ is the bounded (Theorem 7.3) determinant of the Jacobian of ψ^{-1}. From (10.11) it follows that

$$\begin{aligned}&\lim_{n\to\infty}\int_{\psi^{-1}(\mathbb{R}^2\setminus\bar{\mathscr{C}})}(\dot{l}^{(n)}+ll^{(n)})\gamma\\&\quad=\lim_{n\to\infty}\int_{\mathbb{R}^2\setminus\bar{\mathscr{C}}}[\dot{l}^{(n)}\circ\psi^{-1}+(l\circ\psi^{-1})(l^{(n)}\circ\psi^{-1})](\gamma\circ\psi^{-1})|J\psi^{-1}|\\&\quad=0.\end{aligned}$$

On the other hand, $\dot{l}^{(n)}+ll^{(n)}$ converges in the L^∞_{loc}-weak* topology on $[0,\infty)\times S_1$ to $\zeta+l^2=\zeta+\lambda_0^2$ almost everywhere, and the lemma follows. □

THEOREM 10.3. *There is a Lipschitz continuous version of D^2u on $\mathbb{R}^2$.*

Proof. For $\theta\in S_1$ and $0\leqq t<T_1(\theta)$, define

$$\lambda(t,\theta)\triangleq\Delta u(\psi(t,\theta)), \tag{10.12}$$

where, of course, we mean the Lipschitz continuous version of Δu inside $\mathscr{C}$ (Corollary 8.4). For $\theta\in S_1$ and $t\geqq T_1(\theta)$, define $\lambda(t,\theta)$ by (10.4), which gives us a Lipschitz function. At $t=T_1(\theta)$, the Lipschitz continuity of λ follows from (10.4), Lemma 6.1, and the equality $|\nabla u|^2=1$ on $\partial\mathscr{C}$. The Lipschitz continuity of ψ^{-1} implies the Lipschitz continuity of $\lambda\circ\psi^{-1}$.

It remains to show that $\lambda\circ\psi^{-1}$ is a version of Δu, or equivalently,

$$\lambda(t,\theta)=\Delta u(\psi(t,\theta)),\quad\text{a.e. }(t,\theta)\in[0,\infty)\times S_1. \tag{10.13}$$

In light of (10.9) and (10.12), we need only show that for almost every $\theta\in S_1$,

$$\lambda(t,\theta)=\lambda_0(t,\theta)\quad\forall t\geqq T_1(\theta). \tag{10.14}$$

But (10.10) shows that for almost every $\theta \in S_1$, the function $t \mapsto \lambda_0(t, \theta)$ is absolutely continuous on $[0, \infty)$; in particular,

$$\begin{aligned} \lambda_0(T_1(\theta), \theta) &= \lim_{t \uparrow T_1(\theta)} \lambda_0(t, \theta) \\ &= \lim_{t \uparrow T_1(\theta)} \Delta u(\psi(t, \theta)) \\ &= \lim_{t \uparrow T_1(\theta)} [u(\psi(t, \theta)) - h(\psi(t, \theta))] \\ &= u(\psi(T_1(\theta), \theta)) - h(\psi(T_1(\theta), \theta)). \end{aligned} \tag{10.15}$$

Equation (10.10) and Lemma 10.2 imply that for almost every $\theta \in S_1$,

$$\dot{\lambda}_0(t, \theta) = -\lambda_0^2(t, \theta), \quad \text{a.e. } t \geqq T_1(\theta). \tag{10.16}$$

Equations (10.15) and (10.16) imply (10.14). □

11. Regularity of the free boundary. In this section we apply known regularity results for free boundaries to show that the boundary of $\mathscr{C}$ is of class $C^{2,\alpha}$ for all $\alpha \in (0, 1)$. In order to apply these results, we recall that $w = |\nabla u|^2$ is a $W^{2,\infty}$ function (Theorem 8.3) which satisfies (see (5.9)) $1 - w \geqq 0$ on $\mathbb{R}^2$ and

$$\tfrac{1}{2}\Delta(1 - w) = H - w \quad \text{on } \mathscr{C}, \tag{11.1}$$

where we recall that $H \triangleq \nabla h \cdot \nabla u - \|D^2 u\|^2$. We shall establish the strict positivity of the forcing term $H - w$ on $\partial\mathscr{C}$. Recall that

$$w - \tfrac{1}{2}\Delta w - H \leqq 0 \quad \text{on } \mathbb{R}^2,$$

and $w = 1$, $\Delta w = 0$ on $\mathbb{R}^2 \setminus \bar{\mathscr{C}}$, so

$$H - w = H - 1 \geqq 0 \quad \text{on } \mathbb{R}^2 \setminus \mathscr{C}. \tag{11.2}$$

LEMMA 11.1. *The function H is locally Lipschitz continuous, and $H > 1$ on $\partial\mathscr{C}$.*

Proof. The local Lipschitz continuity of H follows from Theorem 10.3. To prove that $H > 1$ on $\partial\mathscr{C}$, we assume that there exists a point on $\partial\mathscr{C}$ where $H = 1$. Without loss of generality, we take this point to be the origin $(0, 0)$, and we take $\nabla u(0, 0) = (-1, 0)$.

We first obtain an upper bound on H near $(0, 0)$. Inside $\mathscr{C}$, H is differentiable and

$$\nabla H \cdot \nabla u = (D^2 h)\nabla u \cdot \nabla u + (D^2 u)\nabla u \cdot \nabla h - \nabla(\|D^2 u\|^2) \cdot \nabla u. \tag{11.3}$$

Let ν^1 and ν^2 be unit eigenvectors for $D^2 u$, and let λ_1 and λ_2 denote their respective (nonnegative) eigenvalues. Then

$$\begin{aligned} \nabla(\|D^2 u\|^2) \cdot \nabla u &= \operatorname{tr}(D^2 w D^2 u) - 2 \operatorname{tr}[(D^2 u)^3] \\ &= \lambda_1 (D^2 w)\nu^1 \cdot \nu^1 + \lambda_2 (D^2 w)\nu^2 \cdot \nu^2 - 2(\lambda_1^3 + \lambda_2^3) \\ &\leqq 2\|D^2 u\|_{L^\infty(\mathscr{C})} \sup_{\nu \in S_1} (D^2 w)\nu \cdot \nu. \end{aligned} \tag{11.4}$$

Applying Theorem 1 and the remark following it from Caffarelli (1977) to the function $1 - w$, we have that for some positive constants C and ε,

$$\sup_{\nu \in S_1} D^2 w(x, y)\nu \cdot \nu \leqq C|\log(\operatorname{dist}((x, y), \partial\mathscr{C}))|^{-\varepsilon} \quad \forall (x, y) \in \mathscr{C}. \tag{11.5}$$

Combining (11.3)–(11.5), we conclude that

$$\begin{aligned} \nabla H(x, y) \cdot \nabla u(x, y) \geqq\ & D^2 h(x, y)\nabla u(x, y) \cdot \nabla u(x, y) + \tfrac{1}{2}\nabla w(x, y) \cdot \nabla h(x, y) \\ & - 2\|D^2 u\|_{L^\infty(\mathscr{C})} C|\log(\operatorname{dist}((x, y), \partial\mathscr{C}))|^{-\varepsilon} \quad \forall (x, y) \in \mathscr{C}. \end{aligned} \tag{11.6}$$

As (x, y) approaches $(0, 0) \in \partial \mathscr{C}$, $|\nabla u(x, y)|$ approaches 1 and $\nabla w(x, y)$ approaches zero. Using (2.8) and (11.6), we can choose $\tilde{\varepsilon} > 0$ such that

$$\nabla H(x, y) \cdot \nabla u(x, y) \geqq \frac{c_0}{2} \quad \forall (x, y) \in [-\tilde{\varepsilon}, \tilde{\varepsilon}]^2 \cap \mathscr{C}. \tag{11.7}$$

Let $\theta_0 \in S_1$ be such that $\psi(T_1(\theta_0), \theta_0) = (0, 0)$. For $t \in (0, T_1(\theta_0))$ chosen so that $\psi(t, \theta_0) \in [-\tilde{\varepsilon}, \varepsilon]^2$,

$$\frac{d}{dt} H(\psi(t, \theta_0)) = \nabla H(\psi(t, \theta_0)) \cdot \nabla u(\psi(t, \theta_0)) \geqq \frac{c_0}{2}.$$

It follows that for some $\tau > 0$,

$$\begin{aligned} H(\psi(T_1(\theta_0) - t, \theta_0)) &\leqq H(\psi(T_1(\theta_0), \theta_0)) - \tfrac{1}{2} c_0 t \\ &= 1 - \tfrac{1}{2} c_0 t \quad \forall t \in (0, \tau). \end{aligned} \tag{11.8}$$

But also

$$\begin{aligned} |\psi(T_1(\theta_0) - t, \theta_0) - (t, 0)| &= |\psi(T_1(\theta_0) - t, \theta_0) - \psi(T_1(\theta_0), \theta_0) \\ &\qquad + t \nabla u(\psi(T_1(\theta_0), \theta_0))| \\ &\leqq t^2 \| D^2 u \|_{L^\infty(\mathscr{C})} \quad \forall t \in (0, T_1(\theta_0)). \end{aligned} \tag{11.9}$$

Let $\beta > 0$ be a Lipschitz constant for H in a sufficiently large neighborhood of $(0, 0)$. From (11.8), (11.9), we have for all $t \in (0, \tau)$:

$$\begin{aligned} H(t, 0) &\leqq H(\psi(T_1(\theta) - t, \theta_0)) + |H(t, 0) - H(\psi(T_1(\theta_0) - t, \theta_0))| \\ &\leqq 1 - \tfrac{1}{2} c_0 t + \beta t^2 \| D^2 u \|_{L^\infty(\mathscr{C})}. \end{aligned}$$

Choosing τ smaller, if necesary, we have $H(t, 0) \leqq 1 - \frac{1}{3} c_0 t$ for all $t \in (0, \tau)$. Again using the Lipschitz continuity of H, we obtain the desired upper bound

$$H(x, y) \leqq 1 - \tfrac{1}{3} c_0 x + \beta |y| \quad \forall (x, y) \in [0, \tau] \times [-\tau, \tau]. \tag{11.10}$$

We next construct a function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ such that for appropriate $\rho, \sigma \in (0, \tau)$,

$$\varphi - \tfrac{1}{2} \Delta \varphi \geqq H \quad \text{on } [0, \rho] \times [-\sigma, \sigma], \tag{11.11}$$

$$\varphi \geqq 1 \quad \text{on } \partial([0, \rho] \times [\sigma, \sigma]), \tag{11.12}$$

$$\varphi(0, 0) = 1. \tag{11.13}$$

For this purpose, choose $0 < \rho < \min\{\tau, (c_0 / 6\sqrt{2}\beta)\}$ such that

$$\left(1 - \frac{\rho^2}{4}\right) \sinh \sqrt{2}\rho \geqq \sqrt{2}\rho. \tag{11.14}$$

Then define

$$\sigma \triangleq \min\left\{\tau, \frac{\rho^2}{4\sqrt{2}}\right\} \tag{11.15}$$

$$A \triangleq \frac{c_0}{3}\left(1 - \frac{\sqrt{2}\rho}{\sinh \sqrt{2}\rho \cosh \sqrt{2}\sigma}\right)^{-1}, \tag{11.16}$$

$$\begin{aligned} \varphi(x, y) \triangleq 1 &+ \beta\sigma\left(2 - \frac{\cosh\sqrt{2}y}{\cosh\sqrt{2}\sigma}\right)\left(1 - \frac{\sinh\sqrt{2}x + \sinh\sqrt{2}(\rho - x)}{\sinh\sqrt{2}\rho}\right) \\ &+ A\rho\left(1 - \frac{\cosh\sqrt{2}y}{\cosh\sqrt{2}\sigma}\right)\left(-\frac{x}{\rho} + \frac{\sinh\sqrt{2}x}{\sinh\sqrt{2}\rho}\right) \quad \forall (x, y) \in \mathbb{R}^2. \end{aligned}$$

Then

$$\varphi(0, y) = \varphi(\rho, y) = 1 \quad \forall y \in [-\sigma, \sigma],$$

$$\varphi(x, \pm\sigma) = 1 + \beta\sigma\left[1 - \frac{\sinh\sqrt{2}x + \sinh\sqrt{2}(\rho - x)}{\sinh\sqrt{2}\rho}\right] \geqq 1 \quad \forall x \in [0, \rho]$$

because

$$\begin{aligned} \sinh a + \sinh b &\leqq \sinh a \cosh b + \sinh b \cosh a \\ &= \sinh(a + b) \quad \forall a, b \in \mathbb{R}. \end{aligned} \tag{11.17}$$

It remains to verify (11.11). Direct computation reveals

$$\begin{aligned} \varphi(x, y) - \frac{1}{2}\Delta\varphi(x, y) &= 1 + 2\beta\sigma - Ax + A\rho\frac{\sinh\sqrt{2}x \cosh\sqrt{2}y}{\sinh\sqrt{2}\rho \cosh\sqrt{2}\sigma} \\ &\quad - \beta\sigma\frac{\cosh\sqrt{2}y}{\cosh\sqrt{2}\sigma}\left(\frac{\sinh\sqrt{2}x + \sinh\sqrt{2}(\rho - x)}{\sinh\sqrt{2}\rho}\right) \\ &\geqq 1 + \beta\sigma - Ax + A\rho\frac{\sqrt{2}x}{\sinh\sqrt{2}\rho \cosh\sqrt{2}\sigma} \\ &\geqq 1 - \left(1 - \frac{\sqrt{2}\rho}{\sinh\sqrt{2}\rho \cosh\sqrt{2}\sigma}\right)Ax + \beta\sigma \\ &\geqq 1 - \tfrac{1}{3}c_0 x + \beta|y| \\ &\geqq H(x, y) \quad \forall (x, y) \in [0, \rho] \times [-\sigma, \sigma], \end{aligned}$$

where we have used (11.17), the inequality $a \leqq \sinh a$ for all $a \geqq 0$, (11.16), and (11.10).

On the other hand, (5.9) implies that

$$\begin{aligned} w - \tfrac{1}{2}\Delta w &\leqq H \quad \text{on } [0, \rho] \times [-\sigma, \sigma] \\ w &\leqq 1 \quad \text{on } \partial([0, \rho] \times [-\sigma, \sigma]). \end{aligned}$$

The maximum principle implies that $w \leqq \varphi$ on $[0, \rho] \times [-\sigma, \sigma]$. In particular, for all $x \in [0, \rho]$,

$$w(x, 0) - w(0, 0) = w(x, 0) - 1 \leqq \sigma(x, 0) - 1 = \varphi(x, 0) - \varphi(0, 0),$$

and thus

$$0 = \frac{\partial}{\partial x} w(0, 0) \leqq \frac{\partial}{\partial x} \varphi(0, 0). \tag{11.18}$$

The final step in the proof is to show that $(\partial/\partial x)\varphi(0, 0) < 0$, so (11.18) is contradicted, as well as the assumption that $H = 1$ at some point on $\partial\mathscr{C}$. We compute

$$\begin{aligned} \frac{\partial}{\partial x}\varphi(0, 0) &= \sqrt{2}\beta\sigma\left(2 - \frac{1}{\cosh\sqrt{2}\sigma}\right)\left(\frac{\cosh\sqrt{2}\rho - 1}{\sinh\sqrt{2}\rho}\right) \\ &\quad - A\left(1 - \frac{1}{\cosh\sqrt{2}\sigma}\right)\left(1 - \frac{\sqrt{2}\rho}{\sinh\sqrt{2}\rho}\right). \end{aligned} \tag{11.19}$$

The first term on the right-hand side of (11.19) is bounded above by

$$2\sqrt{2}\beta\sigma\left(\frac{\cosh\sqrt{2}\rho - 1}{\sinh\sqrt{2}\rho}\right) \leqq 2\beta\sigma\rho.$$

As for the second term, (11.14) and the inequality $\cosh\sqrt{2}\sigma - 1 \geqq \sqrt{2}\sigma$ imply that

$$\begin{aligned} & A\left(1-\frac{1}{\cosh\sqrt{2}\sigma}\right)\left(1-\frac{\sqrt{2}\rho}{\sinh\sqrt{2}\rho}\right) \\ &= \frac{c_0}{3}\left[\left(1-\frac{\sqrt{2}\rho}{\sinh\sqrt{2}\rho}\right)^{-1} + (\cosh\sqrt{2}\sigma - 1)^{-1}\right]^{-1} \\ &\geqq \frac{c_0}{3}\left[\frac{4}{\rho^2}+\frac{1}{\sqrt{2}\sigma}\right]^{-1} \\ &= \frac{c_0}{3}\left(\frac{\sqrt{2}\rho^2\sigma}{\rho^2+4\sqrt{2}\sigma}\right). \end{aligned}$$

Therefore,

$$\frac{\partial}{\partial x}\varphi(0,0) \geqq \sigma\left[2\beta\rho - \frac{c_0}{3}\left(\frac{\sqrt{2}\rho^2}{\rho^2+4\sqrt{2}\sigma}\right)\right],$$

and (11.15) and the choice of ρ show that

$$\frac{\partial}{\partial x}\varphi(0,0) \geqq \sigma\left[2\beta\rho - \frac{c_0}{3\sqrt{2}}\right] < 0. \qquad \square$$

THEOREM 11.2. *The free boundary $\partial\mathscr{C}$ is of class C^1, and w has continuous second partial derivatives inside $\mathscr{C}$ up to $\partial\mathscr{C}$.*

Proof. Because T_1 is Lipschitz (Theorem 9.2), for every $\bar{\theta} \in S_1$, the point $(T_1(\bar{\theta}), \theta)$ is a point of positive density with respect to the measure of Remark 7.4 for the set $\{(t,\theta) \mid \theta \in S^1,\ t \in (T_1(\theta), \infty)\} = \psi(\mathbb{R}^2 \setminus \bar{\mathscr{C}})$. But ψ and ψ^{-1} are locally Lipschitz, so every point of $\partial\mathscr{C}$ is a point of positive Lebesgue density for $\mathbb{R}^2 \setminus \mathscr{C}$. It follows from Theorem 2 of Caffarelli (1977) that $\partial\mathscr{C}$ is Lipschitz. Caffarelli's Theorem 3 can now be applied (with v in Caffarelli's Assumption (H1) equal to our $1-w$), and it yields the desired results. $\square$

COROLLARY 11.3. *The boundary $\partial\mathscr{C}$ is of class $C^{2,\alpha}$ for every $\alpha \in (0,1)$.*

Proof. In light of Theorems 6.3 and 11.2 and equation (6.4), D^2u has a C^1 extension from $\mathscr{C}$ to $\bar{\mathscr{C}}$. Therefore, $H-w$ appearing on the right-hand side of (11.1) has a C^1 extension from $\mathscr{C}$ to $\bar{\mathscr{C}}$, and because $\partial\mathscr{C}$ is of class C^1, $H-w$ has a C^1 extension to an open set containing $\bar{\mathscr{C}}$. (In Lemma 12.4, we explain in some detail how to construct a similar extension.) Lemma 11.1 and Theorem 11.2 permit us to apply a theorem of Kinderlehrer & Nirenberg (1977) (see also Friedman (1982, Thm. 1.1(i), p. 129)), to conclude that $\partial\mathscr{C}$ is of class $C^{1,\alpha}$ for every $\alpha \in (0,1)$.

Now observe that ∇w solves the problem

$$\nabla w - \tfrac{1}{2}\Delta\nabla w = \nabla H \quad \text{in } \mathscr{C},$$

$$\nabla w = 0 \quad \text{on } \partial\mathscr{C}.$$

Since ∇H is continuous up to $\partial\mathscr{C}$ and $\partial\mathscr{C}$ is $C^{1,\alpha}$, Theorem 8.34 of Gilbarg and Trudinger (1983, p. 211), implies that ∇w is of class $C^{1,\alpha}$ on $\mathscr{C}$ up to $\partial\mathscr{C}$. Inserting this regularity into (6.4), we conclude that D^2u, and hence $H-w$, are of class $C^{1,\alpha}$ on $\mathscr{C}$ up to $\partial\mathscr{C}$. We may again appeal to Friedman (1982, Thm. 1.1) to conclude that $\partial\mathscr{C}$ is of class $C^{2,\alpha}$ for every $\alpha \in (0,1)$. $\square$

Remark 11.4. The bootstrapping in Corollary 11.3 can be continued until the regularity of h is exhausted. If, in place of assumption (2.5), we assume that $h \in C^{k,\alpha}_{loc}$ for some $k \geqq 3$ and $\alpha \in (0,1)$, then the free boundary is of class $C^{k,\alpha}$, w is of class $C^{k,\alpha}$

inside $\mathscr{C}$ up to $\partial\mathscr{C}$, and u is of class $C^{k+1,\alpha}$ inside $\mathscr{C}$ up to $\partial\mathscr{C}$. This argument uses Ladyzhenskaya and Ural'tseva (1968, Thm. 1.1, p. 107), to wit, if ∇H is of class $C^{k-3,\alpha}$ up to $\partial\mathscr{C}$ and $\partial\mathscr{C}$ is $C^{k-1,\alpha}$, then ∇w is of class $C^{k-1,\alpha}$ up to $\partial\mathscr{C}$.

12. Construction of the optimal control process.

DEFINITION 12.1. Let $x \in \bar{\mathscr{C}}$ be given. A control process pair $\{(N_t, \zeta_t);\ 0 \leqq t < \infty\}$ as in § 2 is called a solution to the *Skorokhod problem for reflected Brownian motion in* $\bar{\mathscr{C}}$ *starting at x and with reflection direction* $-\nabla u$ *along* $\partial\mathscr{C}$ provided that:

(a) ζ is continuous,

(b) the process X defined by (2.1) satisfies $X_t \in \bar{\mathscr{C}}$, $0 \leqq t < \infty$, almost surely and

(c) for all $0 \leqq t < \infty$,

$$\zeta_t = \int_0^t 1_{\{X_s \in \partial\mathscr{C},\, N_s = -\nabla u(X_s)\}}\, d\zeta_s, \tag{12.1}$$

For every $x \in \bar{\mathscr{C}}$, the Skorokhod problem of Definition 12.1 has a solution starting at x. This follows from Lions and Sznitman (1984, Thm. 4.3), provided that the following three conditions are satisfied:

(C1) $\mathscr{C}$ has a C^1 boundary and satisfies a uniform exterior sphere condition,

(C2) There exists $\sigma > 0$ such that $\nabla u(x) \cdot n(x) \geqq \sigma$ for all $x \in \partial\mathscr{C}$, where $n(x)$ is the outward normal vector for $\mathscr{C}$ at x,

(C3) ∇u on $\mathscr{C}$ has an extension to a C^2 function on an open set containing $\bar{\mathscr{C}}$.

Condition (C1) is implied by Corollary 11.3. We establish (C2) and (C3).

LEMMA 12.2. *Condition* (C2) *is satisfied.*

Proof. Let $x \in \partial\mathscr{C}$ be given. We construct a sequence $\{x_k\}_{k=2}^{\infty}$ in $\mathscr{C}$ such that $x_k \to x$ and $(\nabla w(x_k)/|\nabla w(x_k)|) \to n(x)$. With K as in Lemma 9.1, we have

$$\frac{\nabla w(x_k) \cdot \nabla u(x_k)}{|\nabla w(x_k)|} \geqq \frac{2}{K},$$

and (C2) follows.

As for the construction of $\{x_k\}_{k=2}^{\infty}$, we choose $r > 0$ such that $B_r(x + rn(x)) \cap \mathscr{C} = \phi$. Define $\bar{x} = x + \frac{1}{2}rn(x)$, so $B_{r/2}(\bar{x}) \cap \mathscr{C} = \phi$ and $x \in \partial B_{r/2}(\bar{x})$. Given $k \geqq 2$, we define $\mathscr{C}_k \triangleq \{x \in \mathbb{R}^2;\ w(x) < 1 - (1/k)\}$. We then translate $B_{r/2}(\bar{x})$ in the $-n(x)$ direction until it touches $\partial\mathscr{C}_k$, i.e., we define

$$\rho_k = \sup\{\rho > 0;\ B_{r/2}(\bar{x} - \rho n(x)) \cap \mathscr{C}_k = \phi\},$$

and we choose $x_k \in \overline{B_{r/2}(\bar{x} - \rho_k n(x))} \cap \partial\mathscr{C}_k$. Then $B_{r/2}(\bar{x} - \rho_k n(x))$ is an exterior sphere for $\partial\mathscr{C}_k$ at x_k, so the outward normal to $\mathscr{C}_k$ at x_k is

$$\frac{\nabla w(x_k)}{|\nabla w(x_k)|} = \frac{\bar{x} - \rho_k n(x) - x_k}{|\bar{x} - \rho_k n(x) - x_k|}.$$

As $k \to \infty$, we have $x_k \to x$ and $\rho_k \to 0$, so $(\nabla w(x_k)/|\nabla w(x_k)|) \to n(x)$. □

LEMMA 12.3. *Condition* (C3) *is satisfied.*

Proof. Given $\varepsilon > 0$, we can find a finite set of open discs $\{B_k\}_{k=1}^n$, each with radius ε, such that $\bar{\mathscr{C}} \subset \cup_{k=1}^n B_k$, and we can find C^∞ functions $\gamma_k : \mathbb{R}^2 \to [0, 1]$ such that $\overline{\operatorname{supp} \gamma_k} \subset B_k$ for every k and $\sum_{k=1}^n \gamma_k = 1$ on $\mathscr{C}$. We can decompose ∇u on $\mathscr{C}$ as $\sum_{k=1}^n \gamma_k \nabla u$, so it suffices to show that each $v_k \triangleq \gamma_k \nabla u$ has a C^2 extension from $B_k \cap \mathscr{C}$ to B_k. For sufficiently small $\varepsilon > 0$, in each B_k there is a C^2 change of coordinates which results in $B_k \cap \bar{\mathscr{C}} \subset \{(x, y) \mid x \leqq 0\}$ and $B_k \setminus \bar{\mathscr{C}} \subset \{(x, y) \mid x > 0\}$. Now v_k has a C^2

extension from $B_k \cap \mathscr{C}$ to $B_k \cap \bar{\mathscr{C}}$ (proof of Corollary 11.3), and taking v_k to be zero on $\{(x, y) \mid x \leqq 0\} \setminus (B_k \cap \bar{\mathscr{C}})$, we have a C^2 function on the closed left half-plane. For $x > 0$, $y \in \mathbb{R}$, define

$$v_k(x, y) = 3v_k(0, y) - 3v_k(-x, y) + v_k(-2x, y).$$

It is easy to check that this extended v_k is C^2 on all of $\mathbb{R}^2$. □

THEOREM 12.4. *Let $x \in \mathbb{R}^2$ be given. If $x \in \bar{\mathscr{C}}$, then the solution to the Skorokhod problem of Definition* 12.1 *is an optimal control process pair for the singular stochastic control problem with initial condition x posed in* § 2. *If $x \notin \bar{\mathscr{C}}$, then there exists a unique pair $(t, \theta) \in [0, \infty) \times S_1$ such that $x = \psi(t, \theta)$. Define $\hat{x} \triangleq \psi(T_1(\theta), \theta)$ and let $(\hat{N}, \hat{\zeta})$ be a solution to the Skorokhod problem starting at $\hat{x}$. Then (N, ζ) is optimal for the control problem with initial condition x, where*

$$N_t \triangleq \begin{cases} -\nabla u(\bar{x}) & \text{if } t = 0, \\ \hat{N}_t & \text{if } t > 0, \end{cases} \tag{12.2}$$

$$\zeta_t \triangleq \begin{cases} 0 & \text{if } t = 0 \\ \hat{\zeta}_t + |x - \hat{x}| & \text{if } t > 0. \end{cases} \tag{12.3}$$

In either case, we have that $u(x) = V(x)$, where u is the solution to the HJB *equation* (3.1) (*see Theorem* 4.6), *and V is the value function for the control problem defined by* (2.10).

Proof. The theorem follows immediately from Theorem 3.1 once we observe that in the case $x \notin \bar{\mathscr{C}}$, Lemma 8.1 implies that for all $s \geqq T_1(\theta)$,

$$\begin{aligned} \nabla u(\psi(s, \theta)) &= \nabla u(\hat{x}) + \int_{T_1(\theta)}^{s} \frac{d}{d\tau} \nabla u(\psi(\tau, \theta))\, d\tau \\ &= \nabla u(\hat{x}) + \int_{T_1(\theta)}^{s} D^2 u(\psi(\tau, \theta)) \nabla u(\psi(\tau, \theta))\, d\tau \\ &= \nabla u(\hat{x}). \end{aligned}$$

Thus, when $x \notin \bar{\mathscr{C}}$, the control process pair (N, ζ) of (12.2), (12.3) causes the state to jump from $X_0 = x$ to $X_{0^+} = \hat{x}$ and $u(x) - u(\hat{x}) = |x - \hat{x}|$. After this initial jump, the state is kept inside $\bar{\mathscr{C}}$ by reflection in the $-\nabla u$ direction along $\partial \mathscr{C}$. □

13. Appendix. Proof of Lemma 4.1. For $\varepsilon \in (0, 1)$, $R > 0$, denote by $u^{\varepsilon, R}$ the solution to

$$u^{\varepsilon, R} - \Delta u^{\varepsilon, R} + \beta_\varepsilon(|\nabla u^{\varepsilon, R}|^2) = h \quad \text{on } B_R(0), \tag{13.1}$$

$$u^{\varepsilon, R} = 0 \quad \text{on } \partial B_R(0). \tag{13.2}$$

The existence of $u^{\varepsilon, R} \in C^2(\overline{B_R(0)})$ follows from Ladyzhenskaya and Ural'tseva (1968, Thm. 4.8.3, p. 301); uniqueness follows from the following lemma.

LEMMA 13.1. *Suppose that φ is a subsolution and ψ is a supersolution to* (13.1). *Then for all $x \in B_R(0)$:*

$$\varphi(x) - \psi(x) \leqq \sup_{y \in \partial B_R(0)} [\varphi(y) - \psi(y)]^+. \tag{13.3}$$

Proof. If $\varphi - \psi$ attains its maximum over $\overline{B_R(0)}$ at an interior point x^*, then $\nabla \varphi(x^*) = \nabla \psi(x^*)$ and $0 \geqq \Delta \varphi(x^*) - \Delta \psi(x^*) = \varphi(x^*) - \psi(x^*)$. □

LEMMA 13.2. *Let $q > 0$ be as in* (2.6). *There exists a constant $C_1 > 0$, independent of ε and R, such that*

$$0 \leqq u^{\varepsilon, R}(x) \leqq C_1(1 + |x|^q) \quad \forall x \in B_R(0). \tag{13.4}$$

Proof. To prove the nonnegativity of $u^{\varepsilon,R}$, take $\varphi \equiv 0$ and $\psi = u^{\varepsilon,R}$ in Lemma 13.1. To obtain the upper bound on $u^{\varepsilon,R}$, take $\varphi = u^{\varepsilon,R}$ and

$$\psi(x) = E\int_0^{\tau_x} e^{-t} h(x+\sqrt{2}W_t)\,dt,$$

where $\tau_x \triangleq \inf\{t \geqq 0; |x+\sqrt{2}W_t| \geqq R\}$. Then $\psi - \Delta\psi = h$ on $B_R(0)$, $\psi = 0$ on $\partial B_R(0)$, and Lemma 13.1 and (2.6) imply that

$$\begin{aligned} u^{\varepsilon,R}(x) &\leqq E\int_0^{\tau_x} e^{-t} h(x+\sqrt{2}W_t)\,dt \\ &\leqq E\int_0^{\infty} e^{-t} h(x+\sqrt{2}W_t)\,dt \\ &\leqq 2^q C_0 E\int_0^{\infty} e^{-t}(|x|^q + |\sqrt{2}W_t|^q)\,dt \\ &\leqq C_1(1+|x|^q) \end{aligned}$$

LEMMA 13.3. *There exist constants $C > 0$ and $p > 0$, independent of ε and R, such that*

$$\max_{x\in\partial B_R(0)} |\nabla u^{\varepsilon,R}(x)| \leqq C(1+R^p) \quad \forall \varepsilon \in (0,1), \quad R > 0. \tag{13.5}$$

Proof. Let N be a positive integer greater than $q/2$, and define $g, B : [0,\infty) \to \mathbb{R}$ by

$$g(r) = \sum_{k=0}^{N} \frac{r^{2k}}{4^k (k!)^2}, \quad B(r) = \sum_{k=0}^{\infty} \frac{r^{2k}}{4^k (k!)^2}.$$

Then

$$g(r) - \frac{1}{r} g'(r) - g''(r) = \frac{r^{2N}}{4^N (N!)^2},$$

and

$$B(r) - \frac{1}{r} B'(r) - B''(r) = 0. \tag{13.6}$$

For $R > 0$, define

$$\begin{aligned} \psi_R(x) &= 2C_0 + C_0 4^N (N!)^2 g(|x|) \\ &\quad - [2C_0 + C_0 4^N (N!)^2 g(R)] \frac{B(|x|)}{B(R)} \quad \forall x \in \mathbb{R}, \end{aligned}$$

so

$$\begin{aligned} \psi_R(x) - \Delta\psi_R(x) &= C_0(2+|x|^{2N}) \geqq h(x) \quad \forall x \in B_R(0), \\ \psi_R(x) &= 0 \quad \forall x \in \partial B_R(0). \end{aligned}$$

It follows from Lemma 13.1 that $u^{\varepsilon,R} \leqq \psi_R$ on $B_R(0)$, and because these functions agree on $B_R(0)$ and because $\nabla u^{\varepsilon,R}$ on $\partial B_R(0)$ must point inward, where $u^{\varepsilon,R}$ is nonnegative, we have

$$|\nabla u^{\varepsilon,R}(x)| \leqq |\nabla \psi_R(x)| \quad \forall x \in \partial B_R(0).$$

But on $\partial B_R(0)$,

$$|\nabla \psi_R(x)| = \left| C_0 4^N (N!)^2 g'(R) - [2C_0 + C_0 4^N (N!)^2 g(R)] \frac{B'(R)}{B(R)} \right|.$$

Equation (13.6) and the nonnegativity of B'' show that

$$0 \leqq B'(r) \leqq rB(r) \quad \forall r > 0,$$

so we may bound the growth of $\max_{x \in \partial B_R(0)} |\nabla \psi_R(x)|$ by a constant times $(1 + R^{2N+1})$. □

LEMMA 13.4. *There exist constants $C > 0$, $p > 0$, $\lambda > 0$, independent of ε and R, such that*

$$|\nabla u^{\varepsilon,R}(x)| \leqq \lambda u^{\varepsilon,R}(x) + C|x|^p + C \quad \forall x \in B_R(0), \quad \varepsilon \in (0,1), \quad R > 0. \tag{13.7}$$

Proof. With $C \geqq 1$ and $p \geqq 2$ satisfying (13.5), and C_0 as in (2.7), define $\lambda \triangleq \max\{2, C_0\}$, $B \triangleq Cp^p + C_0$, and consider the auxiliary function

$$\varphi(x) \triangleq \nabla u^{\varepsilon,R}(x) \cdot \nu - \lambda u^{\varepsilon,R}(x) - C|x|^p - B,$$

where $\varepsilon \in (0,1)$, $R > 0$ are fixed, and ν is a fixed unit vector. It suffices to show that $\varphi(x) \leqq 0$ for all $x \in B_R(0)$, so let x^* be a point at which φ attains its maximum over $\overline{B_R(0)}$. If $x^* \in \partial B_R(0)$, then (13.5) implies that $\varphi(x^*) \leqq 0$. Thus, we need only consider the case that $x^* \in B_R(0)$, for which we have

$$0 \geqq \Delta\varphi(x^*) = \Delta \nabla u^{\varepsilon,R}(x^*) \cdot \nu - \lambda \Delta u^{\varepsilon,R}(x^*) - Cp^2 |x^*|^{p-2}.$$

Using (13.1), we may rewrite this as

$$\begin{aligned} 0 \geqq{}& \nabla u^{\varepsilon,R}(x^*) \cdot \nu + 2\beta_\varepsilon'(r^*) \nabla[\nabla u^{\varepsilon,R}(x^*) \cdot \nu] \cdot \nabla u^{\varepsilon,R}(x^*) \\ & - \nabla h(x^*) \cdot \nu - \lambda u^{\varepsilon,R}(x^*) - \lambda \beta_\varepsilon(r^*) + \lambda h(x^*) - Cp^2 |x^*|^{p-2}, \end{aligned} \tag{13.8}$$

where r^* denotes $|\nabla u^{\varepsilon,R}(x^*)|^2$. Because of (2.7),

$$|\nabla h(x)| \leqq C_0 + \lambda h(x) \quad \forall x \in \mathbb{R}.$$

Furthermore,

$$\begin{aligned} Cp^2 |x|^{p-2} &\leqq Cp^p \left| \frac{x}{p} \right|^{p-2} \\ &\leqq C|x|^p + Cp^p \quad \forall x \in \mathbb{R}. \end{aligned}$$

Adding these two inequalities, we see that

$$|\nabla h(x^*)| + Cp^2 |x^*|^{p-2} \leqq \lambda h(x^*) + C|x^*|^p + B.$$

Substitution into (13.8) yields

$$0 \geqq \varphi(x^*) + 2\beta_\varepsilon'(r^*) \nabla[\nabla u^{\varepsilon,R}(x^*) \cdot \nu] \cdot \nabla u^{\varepsilon,R}(x^*) - \lambda \beta_\varepsilon(r^*). \tag{13.9}$$

Because $\nabla \varphi(x^*) = 0$, we also have

$$\begin{aligned} 0 &= \nabla \varphi(x^*) \cdot \nabla u^{\varepsilon,R}(x^*) \\ &= \nabla[\nabla u^{\varepsilon,R}(x^*) \cdot \nu] \cdot \nabla u^{\varepsilon,R}(x^*) - \lambda r^* \\ &\quad - Cp|x^*|^{p-2} x^* \cdot \nabla u^{\varepsilon,R}(x^*). \end{aligned} \tag{13.10}$$

Substitution of (13.10) into (13.9) results in the inequality

$$\varphi(x^*) \leqq \lambda[\beta_\varepsilon(r^*) - 2\beta_\varepsilon'(r^*) r^*] - 2Cp|x^*|^{p-2} \beta_\varepsilon'(r^*) x^* \cdot \nabla u^{\varepsilon,R}(x^*).$$

Let us assume that $\varphi(x^*) > 0$. Then

$$\sqrt{r^*} \geq \nabla u^{\varepsilon,R}(x^*) \cdot \nu \geqq B \geqq 2,$$

so $r^* \geqq 4$ and for all $\varepsilon \in (0,1)$,

$$\beta_\varepsilon(r^*) = \frac{r^*-1}{\varepsilon} - 1, \qquad \beta'_\varepsilon(r^*) = \frac{1}{\varepsilon}.$$

Consequently,

$$0 < \varphi(x^*) \leqq -\frac{\lambda}{\varepsilon}(|\nabla u^{\varepsilon,R}(x^*)|^2 + 1 + \varepsilon) - \frac{2Cp}{\varepsilon}|x^*|^{p-2}x^* \cdot \nabla u^{\varepsilon,R}(x^*)$$

$$\leqq -\frac{\lambda}{\varepsilon}(|\nabla u^{\varepsilon,R}(x^*)|^2 + 1 + \varepsilon) + \frac{2Cp}{\varepsilon}|x^*|^{p-1}|\nabla u^{\varepsilon,R}(x^*)|,$$

which implies that

$$|\nabla u^{\varepsilon,R}(x^*)| \leqq \frac{2Cp}{\lambda}|x^*|^{p-1} \leqq Cp^p \left|\frac{x^*}{p}\right|^{p-1} \leqq C|x^*|^p + B.$$

This inequality contradicts the assumption that $\varphi(x^*) > 0$. □

LEMMA 13.5. *For each $\varepsilon \in (0,1)$, there is an increasing sequence $\{R_n\}_{n=1}^\infty$ of positive numbers converging to infinity and a function $u^\varepsilon \in C^2(\mathbb{R}^2)$ such that $\{u^{\varepsilon,R_n}\}_{n=1}^\infty$ and $\{\nabla u^{\varepsilon,R_n}\}_{n=1}^\infty$ converge uniformly to u^ε and ∇u^ε, respectively, on compact sets. Furthermore, u^ε is a solution to* (4.3) *and satisfies* (4.4), (4.5), *with C_1 and p independent of ε.*

Proof. Let $\varepsilon \in (0,1)$ be fixed and let $r > 0$ be given. Then $u^{\varepsilon,R}$ and $\nabla u^{\varepsilon,R}$ are bounded on $B_{2r}(0)$, uniformly in R and ε (Lemmas 13.2, 13.4). Elliptic regularity implies Hölder continuity of $\nabla u^{\varepsilon,R}$ on $B_r(0)$, uniformly in $R \in [2r, \infty)$ (Gilbarg and Trudinger, Thm. 3.9, p. 41), and by the Arzela–Ascoli Theorem, we can find a sequence $\{R_n\}_{n=1}^\infty$ along which $\{u^{\varepsilon,R_n}\}_{n=1}^\infty$ and $\{\nabla u^{\varepsilon,R_n}\}_{n=1}^\infty$ converge uniformly on $B_r(0)$. Indeed, by diagonalization we can select $\{R_n\}_{n=1}^\infty$ so that $\{u^{\varepsilon,R_n}\}_{n=1}^\infty$ and $\{\nabla u^{\varepsilon,R_n}\}_{n=1}^\infty$ converge uniformly on compact sets to limits u^ε and ∇u^ε, respectively, where $u^\varepsilon \in C^{1,\alpha}$ for all $\alpha \in (0,1)$. Passing to the limit in (13.1), we see that Δu^ε exists in the distributional sense and is equal to $u^\varepsilon + \beta_\varepsilon(|\nabla u^\varepsilon|^2) - h$, which is a $C^{0,\alpha}$ function. Elliptic regularity implies that $D^2 u^\varepsilon$ in fact exists in the classical sense and u^ε is $C^{2,\alpha}$. (By bootstrapping, we could conclude that u^ε is $C^{4,\alpha}$ because h is $C^{2,1}$.) □

The convexity of u^ε will be established by representing u^ε as the value function of a stochastic control problem with convex cost functions. With β_ε defined by (4.2), we define a convex function $g_\varepsilon : \mathbb{R}^2 \to \mathbb{R}$ and its (convex) Legendre transform $l_\varepsilon : \mathbb{R}^2 \to \mathbb{R}$ by

$$g_\varepsilon(x) \triangleq \beta_\varepsilon(|x|^2), \qquad l_\varepsilon(y) \triangleq \sup_{x \in \mathbb{R}^2}\{x \cdot y - g_\varepsilon(x)\}. \tag{13.11}$$

For every $y \in \mathbb{R}^2$,

$$l_\varepsilon(y) \geqq \frac{\varepsilon}{2}|y|^2 - g_\varepsilon\left(\frac{\varepsilon}{2}y\right) \geqq \frac{\varepsilon}{4}|y|^2. \tag{13.12}$$

Furthermore, the supremum in the definition of l_ε is attained if x is related to y by $y = 2\beta'_\varepsilon(|x|^2)\, x$, i.e.,

$$l_\varepsilon(2\beta'_\varepsilon(|x|^2)x) = 2\beta'_\varepsilon(|x|^2)|x|^2 - \beta_\varepsilon(|x|^2) \quad \forall x \in \mathbb{R}^2. \tag{13.13}$$

A *control process* is any two-dimensional, absolutely continuous process η adapted to the Brownian motion $\{W_t, \mathcal{F}_t; 0 \leqq t < \infty\}$ and satisfying $\eta_0 = 0$ almost surely. Given an initial state $x \in \mathbb{R}^2$, the corresponding *state process* is

$$Y_t \triangleq x + \sqrt{2} W_t - \eta_t. \tag{13.14}$$

For each $R > 0$, we define the cost corresponding to η up to the exit from $B_R(0)$ as

$$v_\eta^{\varepsilon,R}(x) \triangleq E^x \int_0^{T_R} e^{-t} [h(Y_t) + l_\varepsilon(\dot{\eta}_t)]\, dt,$$

where $\tau_R \triangleq \inf\{t \geqq 0; |Y_t| \geqq R\}$, and $\dot{\eta}_t = (d/dt)\eta_t$. The value function up to the exit from $B_R(0)$ is

$$v^{\varepsilon,R}(x) \triangleq \inf_\eta v_\eta^{\varepsilon,R}(x).$$

It is clear that $v^{\varepsilon,R}(x)$ is nondecreasing in R, and

$$\lim_{R\to\infty} v^{\varepsilon,R}(x) \leqq v^\varepsilon(x) \triangleq \inf_\eta E^x \int_0^\infty e^{-t} [h(Y_t) + l_\varepsilon(\dot{\eta}_t)]\, dt, \tag{13.15}$$

where v^ε is the value function for a control problem on $\mathbb{R}^2$.

LEMMA 13.6. *For each $\varepsilon \in (0,1)$, $R > 0$, the solution $u^{\varepsilon,R}$ of* (13.1), (13.2) *agrees with $v^{\varepsilon,R}$ on $B_R(0)$.*

Proof. Itô's lemma implies that for a given control process η, $x \in B_R(0)$ and $t \geqq 0$:

$$\begin{aligned} E^x e^{-t\wedge\tau_R} u^{\varepsilon,R}(Y_{t\wedge\tau_R}) &= u^{\varepsilon,R}(x) + E^x \int_0^{t\wedge\tau_R} e^{-s} [\beta_\varepsilon(|\nabla u^{\varepsilon,R}(Y_s)|^2) \\ &\qquad - h(Y_s) - \nabla u^{\varepsilon,R}(Y_s)\cdot \eta_s)]\, ds \\ &\geqq u^{\varepsilon,R}(x) - E^x \int_0^{t\wedge\tau_R} e^{-s} [h(Y_s) + l_\varepsilon(\dot{\eta}_s)]\, ds. \end{aligned} \tag{13.16}$$

Letting $t \to \infty$, we see that $v_\eta^{\varepsilon,R}(x) \geqq u^{\varepsilon,r}(x)$ for all all η, so $v^{\varepsilon,R}(x) \geqq u^{\varepsilon,R}(x)$. However, if Y^R is the solution to

$$Y_t^R = x - \int_0^t 2\beta_\varepsilon'(|\nabla u^{\varepsilon,R}(Y_s^R)|^2) \nabla u^{\varepsilon,R}(Y_s^R)\, ds + \sqrt{2} W_t, \qquad 0 \leqq t \leqq \tau_R,$$

then the corresponding control process satisfies

$$\dot{\eta}_t^R = 2\beta_\varepsilon'(|\nabla u^{\varepsilon,R}(Y_t^R)|^2) \nabla u^{\varepsilon,R}(Y_t^R), \qquad 0 \leqq t \leqq \tau_R,$$

and equality holds in (13.16) because of (13.13), i.e.,

$$v_{\eta^R}^{\varepsilon,R}(x) = u^{\varepsilon,R}(x) \leqq v^{\varepsilon,R}(x),$$

and thus $u^{\varepsilon,R}(x) = v^{\varepsilon,R}(x)$. □

LEMMA 13.7. *For each $\varepsilon \in (0,1)$, the function u^ε constructed in Lemma* 13.5 *agrees with the value function v^ε defined in* (13.15).

Proof. We have immediately from (13.15) and Lemma 13.6 that $u^\varepsilon \leqq v^\varepsilon$. For the reverse inequality, let $x \in \mathbb{R}^2$ be given and define Y^∞ (up to the time of a possible explosion) by

$$Y_t^\infty = x - \int_0^t 2\beta_\varepsilon'(|\nabla u^\varepsilon(Y_t^\infty)|^2) \nabla u^\varepsilon(Y_t^\infty)\, dt + \sqrt{2} W_t.$$

Imitating (13.16), we have from Itô's lemma and (13.13) that for every $R > 0$,

$$u^\varepsilon(x) = E^x \int_0^{t\wedge\tau_R} e^{-s}[h(Y_s^\infty) + l_\varepsilon(\dot\eta_s^\infty)]\,ds + E^x\, e^{-t\wedge\tau_R} u^\varepsilon(Y_{t\wedge\tau_R}^\infty), \tag{13.17}$$

where

$$\dot\eta_t^\infty \triangleq 2\beta_\varepsilon'(|\nabla u^\varepsilon(Y_t^\infty)|^2)\nabla u^\varepsilon(Y_t^\infty), \qquad \tau_R \triangleq \inf\{t \geqq 0; |Y_t^\infty| \geqq R\}.$$

Deleting the (nonnegative) second term on the right-hand side of (13.17) and letting $R \to \infty$, $t \to \infty$, we obtain

$$u^\varepsilon(x) \geqq E^x \int_0^{\tau_\infty} e^{-s}[h(Y_s^\infty) + l_\varepsilon(\dot\eta_s^\infty)]\,ds, \tag{13.18}$$

where $\tau_\infty \triangleq \lim_{R\to\infty} \tau_R$ is finite if and only if Y^∞ explodes in finite time.

To see that $\tau_\infty = \infty$ almost surely, observe that for all $t \geqq 0$, $R > 0$,

$$|\eta_{t\wedge\tau_R}^\infty|^2 = 2\int_0^{t\wedge\tau_R} \eta_s^\infty \cdot \dot\eta_s^\infty\,ds \leqq \int_0^{t\wedge\tau_R} |\dot\eta_s^\infty|^2\,ds + \int_0^{t\wedge\tau_R} |\eta_{s\wedge\tau_R}^\infty|^2\,ds.$$

Gronwall's inequality implies

$$\max_{0\leqq s\leqq t\wedge\tau_R} |\eta_s^\infty|^2 \leqq e^t \int_0^{t\wedge\tau_R} |\dot\eta_s^\infty|^2\,ds \leqq \frac{4e^t}{\varepsilon}\int_0^{t\wedge\tau_R} l_\varepsilon(\dot\eta_s^\infty)\,ds,$$

where we have used (13.12). Letting $R \to \infty$ and taking expectations, we conclude that

$$E^x \sup_{0\leqq s<t\wedge\tau_\infty} |\eta_s^\infty|^2 \leqq E^x \frac{4e^t}{\varepsilon}\int_0^{t\wedge\tau_\infty} l_\varepsilon(\dot\eta_s^\infty)\,ds \leqq \frac{4e^{2t}}{\varepsilon} u^\varepsilon(x) < \infty, \quad \forall t \geqq 0.$$

But

$$\sup_{0\leqq s<t\wedge\tau_\infty} |Y_s^\infty| \leqq x + \sup_{0\leqq s<t\wedge\tau_\infty} |\eta_s^\infty| + \sqrt{2}\max_{0\leqq s\leqq t} |W_s|$$

and $\sup_{0\leqq s<t\wedge\tau_\infty} |Y_s^\infty| < \infty$ on $\{\tau_\infty \leqq t\}$. It follows that $P^*\{\tau_\infty \leqq t\} = 0$ for all $t \geqq 0$. Inequality (13.18) can now be restated as

$$u^\varepsilon(x) \geqq E^x \int_0^\infty e^{-s}[h(Y_s^\infty) + l_\varepsilon(\dot\eta_s^\infty)]\,ds \geqq v^\varepsilon(x). \qquad \square$$

COROLLARY 13.8. *For each* $\varepsilon \in (0, 1)$, *the function* u^ε *constructed in Lemma* 13.5 *is convex.*

COROLLARY 13.9. *For each* $\varepsilon \in (0, 1]$, $\lim_{|x|\to\infty} u^\varepsilon(x) = \infty$.

Proof. In light of (2.8), (2.9), (13.12), and (13.15), we have

$$u^\varepsilon(x) \geqq \inf_\eta E^x \int_0^\infty e^{-t}\left[\frac{c_0}{2}|Y_t|^2 + \frac{\varepsilon}{4}|\dot\eta_t|^2\right] dt.$$

But the right-hand side is the value associated with a linear-quadratic-Gaussian problem, which is easily computed to be $\frac{1}{2}\alpha|x|^2 + 2\alpha$, where α is the positive root of the quadratic equation $(2/\varepsilon)\alpha^2 + \alpha - c_0 = 0$. $\square$

LEMMA 13.10. *There is a constant* C_2, independent of ε, such that for every $\varepsilon \in (0, 1)$, *the function* u^ε *constructed in Lemma* 13.5 *satisfies* (4.6).

Proof. Let ν be a unit vector and define $u_{\nu\nu}^\varepsilon \triangleq (D^2u)\nu \cdot \nu$. It suffices to produce a constant C_2, independent of ε and ν, such that

$$u_{\nu\nu}^\varepsilon \leqq C_2(1 + u^\varepsilon).$$

We begin by differentiating (4.3) to obtain

$$\begin{aligned} h_{\nu\nu} &= u^\varepsilon_{\nu\nu} - \Delta u^\varepsilon_{\nu\nu} + 2\beta'_\varepsilon(|\nabla u^\varepsilon|^2)(\nabla u^\varepsilon_{\nu\nu}\cdot\nabla u^\varepsilon + |(D^2u^\varepsilon)\nu|^2) \\ &\quad + 4\beta''_\varepsilon(|\nabla u^\varepsilon|^2)(D^2u^\varepsilon\nabla u^\varepsilon\cdot\nu)^2 \\ &\geqq u^\varepsilon_{\nu\nu} - \Delta u^\varepsilon_{\nu\nu} + 2\beta'_\varepsilon(|\nabla u^\varepsilon|^2)\nabla u^\varepsilon_{\nu\nu}\cdot\nabla u^\varepsilon. \end{aligned} \tag{13.19}$$

Let x^ε be a minimizing point for u^ε, choose $p>0$ satisfying (4.4), (4.5), choose $C_0>0$ to satisfy (2.8), let $\delta>0$ be given, and define the auxiliary function

$$\varphi_\delta(x) = u^\varepsilon_{\nu\nu}(x) - C_0u^\varepsilon(x) - \delta|x-x^\varepsilon|^{p+2}.$$

This function attains its maximum at some point y^δ, where we have

$$0 = \nabla\varphi_\delta(y^\delta) = \nabla u^\varepsilon_{\nu\nu}(y^\delta) - C_0\nabla u^\varepsilon(y^\delta) - \delta(p+2)|y^\delta-x^\varepsilon|^p(y^\delta-x^\varepsilon), \tag{13.20}$$

$$0 \geqq \Delta\varphi_\delta(y^\delta) = \Delta u^\varepsilon_{\nu\nu}(y^\delta) - C_0\Delta u^\varepsilon(y^\delta) - \delta(p+2)^2|y^\delta-x^\varepsilon|^p. \tag{13.21}$$

Substituting (4.3) into (13.21) and using (13.19), we obtain

$$\begin{aligned} 0 &\geqq u^\varepsilon_{\nu\nu}(y^\delta) + 2\beta'_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2)\nabla u^\varepsilon_{\nu\nu}(y^\delta)\cdot\nabla u^\varepsilon(y^\delta) \\ &\quad - h_{\nu\nu}(y^\delta) - C_0u^\varepsilon(y^\delta) - C_0\beta_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2) \\ &\quad + C_0h(y^\delta) - \delta(p+2)^2|y^\delta-x^\varepsilon|^p \\ &= \varphi_\delta(y^\delta) + 2\beta'_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2)\nabla u^\varepsilon_{\nu\nu}(y^\delta)\cdot\nabla u^\varepsilon(y^\delta) \\ &\quad - h_{\nu\nu}(y^\delta) - C_0\beta_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2) + C_0h(y^\delta) \\ &\quad - \delta(p+2)^2|y^\delta-x^\varepsilon|^p + \delta|y^\delta-x^\varepsilon|^{p+2} \\ &\geqq \varphi_\delta(y^\delta) + 2\beta'_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2)\nabla u^\varepsilon_{\nu\nu}(y^\delta)\cdot\nabla u^\varepsilon(y^\delta) \\ &\quad - C_0(1+h(y^\delta)) - C_0\beta_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2) + C_0h(y^\delta) \\ &\quad - 2\delta p^{(p/2)}(p+2)^{(p+2)/2} \end{aligned} \tag{13.22}$$

because of (2.8) and the fact that

$$-\delta(p+2)^2r^p + \delta r^{p+2} \geqq -2\delta p^{(p/2)}(p+2)^{(p+2)/2} \quad \forall r\geqq 0.$$

But (13.20) implies that

$$\begin{aligned} \nabla u^\varepsilon_{\nu\nu}(y^\delta)\cdot\nabla u^\varepsilon(y^\delta) &= C_0|\nabla u^\varepsilon(y^\delta)|^2 + \delta(p+2)|y^\delta-x^\varepsilon|^p(y^\delta-x^\varepsilon)\cdot\nabla u^\varepsilon(y^\delta) \\ &\geqq C_0|\nabla u^\varepsilon(y^\delta)|^2 \end{aligned} \tag{13.23}$$

because u^ε is convex and attains its minimum at x^ε. Substitution of (13.23) into (13.22) yields

$$\begin{aligned} 0 &\geqq \varphi_\delta(y^\delta) + 2C_0\beta'_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2)|\nabla u^\varepsilon(y^\delta)|^2 - C_0\beta_\varepsilon(|\nabla u^\varepsilon(y^\delta)|^2) \\ &\quad - C_0 - 2\delta p^{(p/2)}(p+2)^{(p+2)/2}. \end{aligned} \tag{13.24}$$

The convexity of β_ε implies that

$$\beta'_\varepsilon(r)r \geqq \beta_\varepsilon(r) - \beta_\varepsilon(0) = \beta_\varepsilon(r) \quad \forall r\geqq 0,$$

so (13.24) reduces to

$$\varphi_\delta(x) \leqq \varphi_\delta(y^\delta) \leqq C_0 + 2\delta p^{(p/2)}(p+2)^{(p+2)/2} \quad \forall x\in\mathbb{R}^2.$$

Letting $\delta\downarrow 0$, we obtain

$$u^\varepsilon_{\nu\nu}(x) \leqq C_0(1+u^\varepsilon(x)) \quad \forall x\in\mathbb{R}^2. \qquad \square$$

REFERENCES

V. E. BENEŠ, L. A. SHEPP, AND H. S. WITSENHAUSEN (1980), *Some solvable stochastic control problems*, Stochastics, 4, pp. 181-207.

J. M. BONY (1967), *Principe du maximum dans les espaces de Sobolev*, C. R. Acad. Sci. Paris, 265, pp. 333-336.

H. BREZIS AND M. SIBONY (1971), *Equivalence de deux inéquations variationelles et applications*, Arch. Rational Mech. Anal., 41, pp. 254-265.

L. A. CAFFARELLI (1977), *The regularity of free boundaries in higher dimensions*, Acta Math., 139, pp. 155-184.

M. CHIPOT (1984), *Variational Inequalities and Flow in Porous Media*, Springer-Verlag, New York.

P.-L. CHOW, J.-L. MENALDI, AND M. ROBIN (1985), *Additive control of stochastic linear systems with finite horizons*, SIAM J. Control Optim., 23, pp. 858-899.

K. DEIMLING (1985), *Nonlinear Functional Analysis*, Springer-Verlag, New York.

G. DUVANT AND H. LANCHON (1967), *Sur la solution du problème de la torsion elasto-plastique d'une barre cylindrique de section quelconque*, C.R. Acad. Sci. Paris Ser. I Math., 264, Série, pp. 520-523.

L. C. EVANS (1979), *A second order elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 4, pp. 555-572. *Erratum*, Ibid., pp. 1199.

A. FRIEDMAN (1982), *Variational Principles and Free Boundary Problems*, John Wiley, New York.

D. GILBARG AND N. TRUDINGER (1983), *Elliptic Partial Differential Equations of Second Order*, Second Edition, Springer-Verlag, New York.

J. M. HARRISON (1985), *Brownian Motion and Stochastic Flow Systems*, John Wiley, New York, 1985.

J. M. HARRISON and A. J. TAYLOR (1978), *Optimal control of a Brownian storage system*, Stochastic Process. Appl., 6, pp. 179-194.

J. M. HARRISON AND M. I. TAKSAR (1983), *Instantaneous control of Brownian motion*, Math. Oper. Res., 8, pp. 454-466.

H. ISHII AND S. KOIKE (1983), *Boundary regularity and uniqueness for an elliptic equation with gradient constraint*, Comm. Partial Differential Equations, 8, pp. 317-346.

I. KARATZAS (1981), *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7, pp. 175-189.

———(1983), *A class of singular stochastic control problems*, Adv. in Appl. Probab., 15, pp. 225-254.

I. KARATZAS AND S. E. SHREVE (1986), *Equivalent models for finite-fuel stochastic control*, Stochastics, 17, pp. 245-276.

——— (1987), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.

D. KINDERLEHRER AND L. NIRENBERG (1977), *Regularity in free boundary problems*, Ann. Scuola Norm. Sup. Pisa Cl. Sci (4), pp. 373-391.

O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA (1986), *Linear and Quasilinear Elliptic Equations*, Academic Press, New York.

J. P. LEHOCZKY AND S. E. SHREVE (1986), *Absolutely continuous and singular stochastic control*, Stochastics, 17, pp. 91-109.

P.-L. LIONS (1983), *A remark on Bony maximum principle*, Proc. Amer. Math. Soc., 88, pp. 503-508.

P.-L. LIONS AND A. S. SZNITMAN (1984), *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37, pp. 511-537.

J.-L. MENALDI AND M. ROBIN (1983), *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278, pp. 771-802.

P. A. MEYER (1976), Lecture Notes in Math. 511, Séminaire de Probabilités X, Université de Strasbourg, Springer-Verlag, New York.

S. E. SHREVE, J. P. LEHOCZKY, AND D. P. GAVER (1984), *Optimal consumption for general diffusions with absorbing and reflecting barriers*, SIAM J. Control Optim., 22, pp. 55-75.

M. SUN (1987), *Singular control problems in bounded intervals*, Stochastics, 21, pp. 303-344.

M. I. TAKSAR (1985), *Average optimal singular control and a related stopping problem*, Math. Oper. Res., 10, pp. 63-81.

T. W. TING (1966), *Elastic-plastic torsion of a square bar*, Trans. Amer. Math. Soc., 123, pp. 369-401.

T. W. TING (1967), *Elastic-plastic torsion problem II*, *Arch. Rational Mech. Anal.*, 25, pp. 342-366.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 909–915, September 1989

A TRIBUTE TO E. J. McSHANE

This special issue of the *SIAM Journal on Control and Optimization* is dedicated to E. J. McShane on the occasion of his 85th birthday.[1] During his long career, he has had a leading role in the development of several mathematical areas that bear significantly on present-day control theory. In addition, he served in several capacities as a national leader in mathematical and science policy matters. Those of us who have had the privilege to know Jim McShane have the highest regard for him as a mathematician, a scientific statesman, and a person.

McShane was born in New Orleans on May 10, 1904, and grew up there. His father was a medical doctor and his mother a former school teacher. He graduated from Tulane University in 1925, receiving simultaneously Bachelor of Engineering and Bachelor of Science degrees. He turned down an offer from General Electric and instead continued as a student instructor of mathematics at Tulane, receiving a Master's degree in 1927.

In the summer of 1927, McShane entered graduate school at the University of Chicago, from which he received the Ph.D. in 1930 under the supervision of G. A. Bliss. He interrupted his studies during 1928–29 for financial reasons to teach at the University of Wichita. It was at Chicago that McShane's long-standing interest in the calculus of variations began. From 1930 to 1932 he held a National Research Council fellowship, spent at Princeton, Ohio State, Harvard, and Chicago. This was a very productive period. It resulted in thirteen research papers, containing a wealth of new ideas. Another fortunate event was his marriage to Virginia Haun in 1931.

Because of the Great Depression, openings in mathematics departments were virtually nonexistent in 1932. The McShanes spent 1932–33 at Gottingen, during which time he translated into English the two volumes of Courant's *Differential and Integral Calculus*. They also saw firsthand some frightening aspects of the onset of Nazi power in Germany.

After two years (1933–35) on the Princeton faculty, McShane joined the Department of Mathematics at the University of Virginia as a full professor in the fall of 1935. He has remained there ever since. In 1939 and 1940, McShane's important papers on general necessary conditions in the form of multiplier rules and on a strong existence theorem for the Bolza problem of calculus of variations appeared. With the onset of World War II, McShane agreed to head a mathematics group at the Ballistics Research Laboratory in Aberdeen, Maryland. During this time he wrote a book with J. L. Kelley and F. V. Reno entitled *Exterior Ballistics*, which is regarded as the definitive work on the subject.

After the war, McShane developed a serious interest in the mathematical foundations of quantum mechanics and quantum field theory. While this ambitious program did not reach fruition, the attempt profoundly influenced his subsequent work on integration processes and stochastic calculus. This is seen, for example, in his excellent *Bulletin of the American Mathematical Society* survey article "Integrals Designed for Special Purposes" (1963) and his book *Stochastic Calculus and Stochastic Models* (1974), which is the definitive treatment of his approach to that subject.

McShane served as President of the Mathematical Association of America during 1953–54. He took an active interest in efforts just then getting underway to revitalize undergraduate mathematics in the U.S. During McShane's term as President, the MAA

[1] After this issue was compiled, we were saddened to learn of the death of E. J. McShane on June 1, 1989, at the age of 85.

Committee on the Undergraduate Program in Mathematics was established. Since that time the Committee has been a leader in these endeavors. He was elected to the National Academy of Science in 1948 and served on the National Science Board from 1956 to 1968. During 1958 and 1959, he was also President of the American Mathematical Society.

McShane has had a lifelong interest in music. His early interest in opera led him to learn to read Italian libretti. This knowledge, in turn, led G. A. Bliss to suggest to McShane that he read the then new book *Fondamenti di Calcolo delle Variazione* by Leonida Tonelli, which started McShane on his study of multiple integral problems in the calculus of variations. Later, in the 1950s, McShane learned to play the cello, and he has been an amateur chamber music performer ever since.

The injustices suffered by some of his colleagues during the post World War II anticommunist hysteria deeply offended McShane. He himself, in response to the question on the Aberdeen Proving Grounds security form that asked whether he had ever been involved with organizations that at any time advocated the overthrow of the U.S. government by force and violence, replied that, yes, he was an employee of the state of Virginia. During the McCarthy era, the House Un-American Activities Committee (HUAC) "invited" him to express his views, but he was not subpoened. He did not cooperate with HUAC, but wrote a letter in which he stated his views and backed them up with quotations from various sources.

Victor Klee, recalling his experience as a graduate student at Virginia from 1945 to 1949, writes: "... He [McShane] was very popular with the graduate students because of his clear lectures, his amusing anecdotes, and unusual kindness." Klee goes on to tell how McShane turned his office over to the graduate students, who had no offices of their own, and says, "... His generosity contributed a lot to the quality of the graduate program by providing a place for the graduate students to meet with each other and talk about mathematics.... It is simply impossible, in a few words, to convey the extent of the graciousness, kindness, and hospitality that have been [and are] exhibited by Virginia and Jimmy McShane in their relations with those lucky enough to know them. These go far beyond professional matters."

The portion of McShane's work that is significant for present-day control theory falls into four broad categories: (a) Multiple integral problems in the calculus of variations; (b) Relaxed controls, necessary conditions and existence theorems for single integral problems and control problems; (c) Integration theory; (d) Stochastic calculus.

(a) The late 1920s and 1930s saw many changes in the calculus of variations. L. Tonelli's book had introduced the "direct method," which was advantageous for proving semicontinuity and the existence of absolute minima. The solution to Plateau's problem by J. Douglas and T. Rado stimulated the rapid development of the calculus of variations for multiple integral problems and the theory of Lebesgue area of surfaces. (The Plateau problem is to find a surface of minimum area with given boundary.) McShane was at the forefront of these developments. While still a graduate student, McShane obtained the necessary condition of Weierstrass for quasiconvex variational problems with an arbitrary number of functions of several variables. Soon afterward he turned to questions of semicontinuity and existence of a minimum for multiple integral geometric calculus of variations, of which the Plateau problem was a prototype. Hidden in these problems were notorious analytical and topological difficulties, which were later overcome by other mathematicians (including Cesari, Federer, and Rado) as part of Lebesgue surface area theory. McShane provided an elegant solution for geometric variational integrands which do not vary spatially. The key idea was that it suffices to find the minimum in the smaller class of "saddle surfaces,"

which are representable parametrically by a vector function monotone in Lebesgue's sense.

(b) In 1939 McShane published a paper in the *American Journal* entitled "On Multipliers for Lagrange Theory," which has had a profound, but not generally recognized, influence on optimal control theory and nonlinear programming. In this paper, McShane showed that the Weierstrass condition holds along a curve, without making any assumptions about the normality of the curve. Although this result was important in and of itself, the method of proof turned out to be a major contribution to the theory of necessary conditions in optimization problems. The key and novel elements of the proof were, first, the construction of a convex cone generated by first-order approximations to the end points of perturbations of the optimal trajectory and, second, showing that optimality implies that this cone and a certain half-ray can be separated by a hyperplane. Twenty years later, this idea was used by Pontryagin and his coworkers in their proof of the necessary condition now known as the Pontryagin maximum principle. In his 1959 Uspekhi paper, Pontryagin states that the proof will utilize certain constructions due to McShane. No such acknowledgment exists in the classic book by Pontryagin, Boltyanskii, Gamkrelidze, and Mischchenko entitled *The Mathematical Theory of Optimal Processes*, which collected their previous work. This book appears to have popularized the convex cone and separation constructions, which were subsequently used by most authors in deriving necessary conditions, not only for control problems, but also for nonlinear programming problems and abstract optimization problems.

Another body of work, which was definitive, in a sense, for problems in the calculus of variations in one independent variable, was the series of three papers that appeared in 1940 in volumes six and seven of the *Duke Journal.* In the first of these papers, McShane showed that if the problem of Bolza is phrased in terms of generalized curves (which were introduced in 1937 for simple problems in the plane by L. C. Young) then the problem of Bolza has a solution. In the second paper, he derived the generalizations of the standard necessary conditions that must hold along a minimizing generalized curve. In the last paper, he gave conditions under which the minimizing generalized curve is an ordinary curve. Definitive as this work was, it did not seem to attract attention outside the circle of cognoscenti in the calculus of variations until twenty years later, in the 1960s, when generalized curves were rediscovered by control theorists as relaxed controls, or sliding states. In 1967, McShane, in a *SIAM Journal on Control* paper, adapted his 1940 work to the control theory setting [*SIAM J. Control.*, 5 (1967), pp. 438–485]. This paper is more elementary and self-contained than most treatments of relaxed controls and reflects McShane's dedication to teaching as well as research.

McShane's proof, with R. B. Warfield, of a general version of Filippov's implicit function theorem (*Proceedings of the American Mathematical Society*, 1967; corrigenda and addenda, 1969) was an important contribution to control theory. This lemma gives conditions that guarantee the existence of a measurable solution to an equation whenever a pointwise solution exists and is one of the basic tools in optimal control theory.

Another example of McShane's interest in instruction is his 1973 paper in the *American Mathematical Monthly* entitled "The Lagrange Multiplier Rule." Here he gives a penalty function proof of the Fritz-John and Kuhn-Tucker necessary conditions for nonlinear programming problems that is short and accessible to anyone who knows the Bolzano-Weierstrass Theorem. Later, other authors applied the arguments used here to obtain necessary conditions for a variety of control and optimization problems.

(c) Over the years McShane has achieved an extraordinarily deep understanding of integration processes as they arise in various guises. He wrote three books on integration, in addition to a number of research articles and the 1963 *Bulletin of the American Mathematical Society* survey already mentioned. His 1944 volume *Integration* gave a readable introduction to the Lebesgue theory at a time when few such books existed in English. The 1953 monograph, *Order Preserving Maps and Integration Processes*, was an outgrowth of his search for a mathematically correct setting in which to treat divergent integrals in quantum physics. In 1957, J. Kurzweil defined a modification of the Riemann integral, which turned out to be more general than the Lebesgue integral. McShane's 1983 volume *Unified Integration* develops in a similar vein a complete theory of integrals, together with a wealth of applications to physics, differential equations, and probability. An appealing feature of this approach, from a pedagogical standpoint, is that point set topology and measurability issues can be deferred.

(d) During the 1960s and 1970s McShane's interests turned toward developing a stochastic differential and integral calculus. The K. Itô stochastic calculus was by then already in existence. It provided a convenient way to represent an important class of stochastic processes, called Markov diffusions, as the solutions to stochastic differential equations. The random inputs to an Itô-sense stochastic differential equation are Brownian motion processes, whose formal time derivatives are "white noises." At that time, however, there was considerable confusion in the engineering literature about the correct interpretation if an idealized white noise is replaced either by a physical "wide band" noise or by some discrete process introduced for numerical approximation to the solution of the stochastic differential equation. This issue was clarified by the work of McShane, Stratonovich, and Wong-Zakai.

McShane's solution was to introduce a stochastic calculus, in which stochastic differential equations take the form (for scalar-valued processes x_t, z_t)

$$dx_t = f(x_t)\,dt + g(x_t)\,dz_t + h(x_t)(dz_t)^2,$$

where z_t is a stochastic process representing the random inputs. If z_t has Lipschitz sample paths, then one should take $(dz_t)^2 = 0$; while $(dz_t)^2 = dt$ for a standard Brownian motion z_t. Let $h = \frac{1}{2}gg'$. Then McShane's stochastic integral has the following consistency property. Let $z_t^{(n)}$ be a sequence of processes with Lipschitz sample paths, such that $z_t^{(n)}$ tends (in a suitable sense) to a Brownian motion z_t on a time interval $0 \leqq t \leqq T$. Then the solution $x_t^{(n)}$ to the $z_t^{(n)}$-driven stochastic differential equation tends to the corresponding solution x_t of the solution to the z_t-driven stochastic differential equation. The function h is called the Wong-Zakai correction. McShane's *Stochastic Calculus and Stochastic Models* (1974) gives a definitive account of this work. Even today the consistency question is often not addressed in the applied literature in such areas as chemical physics, financial economics, and biology. Consistency becomes a more delicate matter when T is large (or infinite) as happens in questions of large deviations or ergodicity. It is perhaps ironic that it has been left to probabilists to sort out these practical consistency questions.

For these many contributions and services to mathematics in general and to control theory in particular, we thank and honor Jim McShane, a true scholar and gentleman.

L. D. Berkovitz
Purdue University

Wendell H. Fleming
Brown University

PUBLICATIONS BY E. J. McSHANE

BOOKS

Integration, Princeton University Press, Princeton, NJ, 1944.
Order-Preserving Maps and Integration Processes, Princeton University Press, Princeton, NJ, 1953.
with J. L. Kelley and F. V. Reno, *Exterior Ballistics*, University of Denver Press, Denver, CO, 1953.
with T. A. Botts, *Real Analysis*, Van Norstrand, New York, 1959.
Stochastic Calculus and Stochastic Models, Academic Press, New York, 1974.
Unified Integration, Academic Press, New York, 1983.

ARTICLES

Semi-continuity in the calculus of variations, and existence theorems for isoperimetric problems, Ph.D. thesis, in Contributions to the Calculus of Variations, University of Chicago Press, Chicago, IL, 1930, pp. 195-243.
On the necessary condition of Weierstrass in the multiple integral problem of the calculus of variations, Ann. of Math. (2), 32 (1930), pp. 578-590.
On the necessary condition of Weierstrass in the multiple integral problem of the calculus of the variations, II, Ann. of Math. (2), 32 (1930), pp. 723-733.
A remark concerning the necessary conditions of Weierstrass, Bull. Amer. Math. Soc., 37 (1931), pp. 631-632.
Remark concerning Mr. Graves paper "On an existence theorem of the calculus of variations," Monatsh. Math., 39 (1931), p. 105.
On a certain inequality of Steiner, Ann. of Math. (2), 33 (1932), pp. 125-138.
On the semi-continuity of double integrals in the calculus of variations, Ann. of Math. (2), 33 (1932), pp. 460-484.
Parametrization of saddle surfaces with application to the problem of Plateau, Trans. Amer. Math. Soc., 35 (1933), pp. 716-733.
Uber die unlosbarkeit eines einfachen problems der variationsrechnung, Nacht. Ges. Wiss. Göttingen I Nr., 45 (1933), pp. 359-364.
Existence theorems for ordinary problems of the calculus of variations, Part I, Ann. Scula Norm. Sup. Pisa. Cl. Sci. (2), 3 (1934), pp. 183-212.
Concerning the semi-continuity of ordinary integrals of the calculus of variations, Ann. Scula Norm. Sup. Pisa. Cl. Sci. (2), 3 (1934), pp. 239-241.
Existence theorems for ordinary problems of the calculus of variations, Part II, Ann. Scula Norm. Sup. Pisa. Cl. Sci. (2), 3 (1934), pp. 287-315.
Integrals over surfaces in parametric form, Ann. of Math. (2), 34 (1933), pp. 815-838.
The DuBois-Reymond relation in the calculus of variations, Math. Ann., 109 (1934), pp. 746-755.
On the analytic nature of surfaces of least area, Ann. of Math. (2), 35 (1934), pp. 456-475.
On the minimizing property of the harmonic function, Bull. Amer. Math. Soc., 40 (1934), pp. 593-598.
Extension of range of functions, Bull. Amer. Math. Soc., 40 (1934), pp. 837-842.
Existence theorems for double integral problems of the calculus of variations, Trans. Amer. Math. Soc., 38 (1935), pp. 459-563.
Semi-continuity of integrals in the calculus of variations, Duke Math. J., 2 (1936), pp. 597-616.
A navigation problem in the calculus of variations, Amer. J. Math., 59 (1937), pp. 327-334.
Jensen's inequality, Bull. Amer. Math. Soc., 43 (1937), pp. 521-527.
On the Osgood-Caratheodory theorem, Amer. Math. Monthly, 44 (1937), pp. 288-291.
Some existence theorems for problems in the calculus of variations, Duke Math. J., 4 (1938), pp. 132-156.
Recent developments in the calculus of variations, Amer. Math. Soc., Semicentennial Publications, vol. II, 1938, pp. 69-97.
Some existence theorems in the calculus of variations. I. *The Dresden corner condition*, Trans. Amer. Math. Soc., 44 (1938), pp. 429-438.
Some existence theorems in the calculus of variations. II. *Existence theorems for isoperimetric problems in the plane*, Trans. Amer. Math. Soc., 44 (1938), pp. 439-453.
Some existence theorems in the calculus of variations. III. *Existence theorems for nonregular problems*, Trans. Amer. Math. Soc., 45 (1939), pp. 151-171.
Some existence theorems in the calculus of variations. IV. *Isoperimetric problems in non-parametric form*, Trans. Amer. Math. Soc., 45 (1939), pp. 173-196.
Some existence theorems in the calculus of variations. V. *The isoperimetric problem in parametric form*, Trans. Amer. Math. Soc., 45 (1939), pp. 197-216.
The Jacobi condition and the index theorem in the calculus of variations, Duke Math. J., 5 (1939), pp. 184-206.
On multipliers for Lagrange problems, Amer. J. Math., 61 (1939), pp. 809-819.
On the uniqueness of the solutions of differential equations, Bull. Amer. Math. Soc., 45 (1939), pp. 755-757.

Curve-space topologies associated with variational problems, Ann. Scula Norm. Sup. Pisa. Cl. Sci. (2), 4 (1940), pp. 45-60.

An estimate of the Weierstrass $\mathscr{E}$-function, Ann. of Math., 41 (1940), pp. 314-320.

with M. R. Hestenes, *A theorem on quadratic forms and its application in the calculus of variations*, Trans. Amer. Math. Soc., 47 (1940), pp. 501-512.

A remark concerning sufficiency theorems for the problem of Bolza, Bull. Amer. Math. Soc., 46 (1940), pp. 698-701.

Generalized curves, Duke Math. J., 6 (1940), pp. 513-536.

Necessary conditions in generalized-curve problems of the calculus of variations, Duke Math. J., 7 (1940), pp. 1-27.

Existence theorems for Bolza problems in the calculus of variations, Duke Math. J., 7 (1940), pp. 28-61.

On the second variation in certain normal problems of the calculus of variations, Amer. J. Math., 63 (1941), pp. 516-530.

Sobre la teoria de los extremos relativos, Revista de Ciencias (Lima), ano 42 (1941), pp. 111-134; ano 43 (1941), pp. 475-482; pp. 659-666; ano 44 (1941), pp. 85-92.

Computation of flat trajectories with high angles of departure, Amer. Math. Monthly, 48 (1941), pp. 617-623.

The addition formulas for the sine and cosine, Amer. Math. Monthly, 48 (1941), pp. 688-689.

On Perron integration, Bull. Amer. Math. Soc., 48 (1942), pp. 718-726.

Sufficient conditions for a weak relative minimum in the problem of Bolza, Trans. Amer. Math. Soc., 52 (1942), pp. 344-379.

An interpolation formula, Amer. Math. Montly, 53 (1946), pp. 259-264.

Review of "Length and Area," by T. Rado, Bull. Amer. Math. Soc., 54 (1948), pp. 861-863.

Remark concerning integration, Proc. Nat. Acad. Sci., 35 (1949), pp. 46-49.

Images of sets satisfying the condition of Baire, Ann. of Math., 51 (1950), pp. 380-386.

The differentials of certain functions in exterior ballistics, Duke Math. J., 17 (1950), pp. 115-134.

Linear functions on certain Banach spaces, Proc. Amer. Math. Soc., 1 (1950), pp. 402-408.

A metric in the space of generalized curves, Ann. of Math., 52 (1950), pp. 328-349.

Partial orderings and Moore-Smith limits, Amer. Math. Monthly, 54 (1952), pp. 1-11.

with T. A. Botts, *A modified Riemann-Stieltjes integral*, Duke Math. J., 19 (1952), pp. 293-302.

The spectrum of the harmonic oscillator, Virginia J. Sci., 4 (1953), pp. 7-10.

The theory of convergence, Canad. J. Math., 6 (1954), pp. 161-168.

A dominated-convergence theorem, Duke Math. J., 22 (1955), pp. 325-332.

On Stieltjes integration, Proc. Amer. Math. Soc., 7 (1956), pp. 69-74.

Maintaining communication, Amer. Math. Monthly, 64 (1957), pp. 309-317.

Operating with sets, reprinted from the 23rd Yearbook of NCTM, (1957), pp. 36-64.

Gilbert Ames Bliss, Biographical Memoirs, Nat. Acad. of Sciences, 31 (1958), pp. 32-53.

A canonical form for antiderivatives, Illinois J. Math., 3 (1959), pp. 334-351.

The Fourier transform and mean convergence, Amer. Math. Monthly, 68 (1961), pp. 205-211.

Curves of constant breadth, Math. Student, 8 (1961), nos. 2 and 3.

A theory of limits, in Studies in Modern Analysis, R. Buck, ed., MAA Studies in Mathematics, vol. 1, Mathematical Association of America, Washington, DC, 1962.

Families of measures and representations of algebras of operators, Trans. Amer. Math. Soc., 102 (1962), pp. 328-345.

Stochastic integrals and non-linear processes, J. Math. Mech., 11 (1962), pp. 235-284.

Weak topologies for stochastic processes, Proc. Nat. Acad. Sci., 48 (1962), pp. 1148-1151.

New ideas in the teaching of mathematics in the colleges of the United States, in Mathematical Education in the Americas—A Report on the first Inter-American Conference on Mathematical Education, Bogota, Columbia, December, 1961, Bureau of Publications, Teachers' College, Columbia, New York, 1962, pp. 97-109. (Also in Spanish.)

The radius of gyration of a convex body, Proc. Amer. Math. Soc., 13 (1962), pp. 922-926.

A weak topology for stochastic processes, Bol. Soc. Mat. Sao Paulo, 14 (1962), pp. 85-101.

Integrals devised for special purposes, Bull. Amer. Math. Soc., 69 (1963), pp. 597-627.

with A. S. Galbraith and G. B. Parrish, *On the solutions of linear second-order differential equations*, Proc. Nat. Acad. Sci., 53 (1965), pp. 247-249.

Integration in linear spaces, Arch. Rational Mech. Anal., 18 (1965), pp. 403-421; *Erratum*, Arch. Rational Mech. Anal., 23 (1967), p. 409.

Why and how the C.U.P. *began*, in Proc. of Preliminary Meetings on College Level Mathematics Education, under the auspices of the U.S.-Japan Program on Scientific Cooperation, Katada, Japan, 1964, Japan Soc. for the Promotion of Sci., Tokyo, 1965, pp. 39-48.

A generalization of convexity, and martingales in linear spaces, Proc. Nat. Acad. Sci., 54 (1965), pp. 37-40.

On the solutions of the differential equation $y' + p^2 y = 0$, Proc. Amer. Math. Soc., 17 (1966), pp. 55-61.
Trends in analysis, Amer. Math. Monthly, 74 (1967), pp. 65-79
Award for distinguished service to Prof. William Larkin Duren, Jr., Amer. Math. Monthly, 74 (1967), pp. 1-2.
with R. B. Warfield, Jr., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc., 18 (1967), pp. 41-47.
Relaxed controls and variational problems, SIAM J. Control, 5 (1967), pp. 438-485.
Optimal controls, relaxed and ordinary, in Math. Theory of Control, A. V. Balakirshnan and L. W. Neustadt, eds., Academic Press, New York, 1967, pp. 1-9.
Optimal control theory and stochastic differential equations, in Proc. U.S.-Japan Seminar on Differential and Functional Equations, University of Minnesota, 1967, W. A. Benjamin, New York, 1967, pp.225-247.
with R. B. Warfield, Jr. and V. M. Warfield, *Invariant extensions of linear functionals, with application to measures and stochastic processes*, Pacific J. Math., 28 (1969), pp. 121-142.
Stochastic integrals and stochastic functional equations, SIAM J. Appl. Math., 17 (1969), pp. 287-306.
A Riemann-type integral that includes Lebesgue-Stieltjes, Bochner, and stochastic integrals, Memoir 88, American Mathematical Society, Providence, 1969.
Addenda and corrigenda to "On Filippov's implicit functions lemma," Proc. Amer. Math. Soc., 21 (1969), pp. 496-498.
Stochastic functional equations: Continuity properties and relation to ordinary differential equations, in Proc. of the Workshop on Calculus of Variations and Optimal Control Theory, II, University of California at Los Angeles, 1968, Academic Press, New York, 1969, pp. 30-103.
Toward a stochastic calculus, I. Proc. Nat. Acad. Sci., 63 (1969), pp. 275-280.
Toward a stochastic calculus, II, Proc. Nat. Acad. Sci., 63 (1969), pp. 1084-1087.
Vector spaces and their applications, in The Mathematical Sciences, the N.R.C.'s Committee on the Support of Research in the Mathematical Sciences, eds., M.I.T. Press, Cambridge, MA, 1969.
On the necessary condition of Weierstrass in the multiple integral problem of the calculus of variations, III, Rend. Circ. Mat. Palermo (2), 26 (1967), pp. 321-345.
Stochastic differential equations and models of random processes, in Proc. VI Berkeley Symposium on Math., Statistics, and Probability, vol. III, University of California Press, Berkeley, CA, 1972, pp. 263-294.
Stochastic models, in Proc. of the Fourth IFIP Colloqium on Optimization Techniques, Los Angeles, CA, 1972, A. V. Balakrishnan, ed., Academic Press, NY, 1972, pp. 311-324.
A unified theory of integration, Amer. Math. Monthly, 80 (1973), pp. 349-359.
Stochastic integration, in Proc. of a Conference on Vector and Operator Measures, Park City, UT, 1972.
The Lagrange multiplier rule, Amer. Math. Monthly, 80 (1973), pp. 922-924.
Stochastic differential equations, J. Multivariate Anal., 5 (1975), pp. 121-177.
Stochastic differential equations and models of noisy systems, in Proc. of 1975 Conference on Information Science and Systems, the Johns Hopkins University, Baltimore, MD, 1975.
The choice of stochastic model for a noisy system, in Proc. of Symposium on Stochastic Optimization, University of Kentucky, 1975.
The calculus of variations from the beginning through optimal control theory, in Proc. of 1977 Oklahoma Conference on Control Theory and Differential Equations, 1977.
with R. B. Darst, *The deterministic Ito-belated integral is equivalent to the Lebesgue integral*, Proc. Amer. Math. Soc., 72 (1978), pp. 271-275.
Choosing a mathematical model for a system affected by noise, Trans. of the 24th Conference of Army Mathematicians, University of Virginia, Charlottesville, VA, 1978, pp. 1-11; ARO Report 79, U.S. Army Research Office, Research Triangle Park, NC, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 916-939, September 1989

THE CALCULUS OF VARIATIONS FROM THE BEGINNING THROUGH OPTIMAL CONTROL THEORY*

E. J. McSHANE†

Before I begin this talk, I would like to sketch briefly what I plan to do. I hope to speak of some of the important stages of the development of the calculus of variations, with a disproportionately large part of the hour allotted to recent developments. But I have no intention of listing important discoveries with their dates. Rather, I shall try to say something of the underlying patterns of thought at each stage and to comment on the change in that pattern produced by each of the new ideas. It may seem that I am deriding our predecessors for not having seen at once all that we have learned. I have no such intention. We must all do our thinking on the foundation of what we already know. It is hard to assimilate a genuinely new idea, and even harder to realize that ideas we have earlier acquired have become obsolete.

Preparing this talk has forced me to formulate with at least some pretension to clarity what is meant by the calculus of variations. There is no universal agreement on the definition of the subject, and I have gradually come to the conclusion that part of the reason is that there are at least two related but different sets of ideas that are often brought together under the same name. The first set might be called the theory of extrema. A functional is defined on some class of functions; the problem is to find a function in the given class that minimizes or maximizes the functional on that class. If this theory of extrema is included in the calculus of variations, Caratheodory may be justified in asserting that the first problem in the calculus of variations was that of finding a curve of given length that joins the ends of a line segment, and together with that segment encloses the greatest possible area. This was solved, according to Caratheodory, by Pappus, in about A.D. 290.

The second set of ideas is concerned with functionals on linear topological spaces, usually function spaces, and constitutes a part of a differential calculus on such spaces. The central problem in this part of the theory is that of finding stationary points of functionals; that is, points at which the directional derivatives in all directions exist and are all zero. Since such points are characterized by means of investigating the effect on the functional produced by small variations of the function which is the independent variable, this study of stationary points can reasonably be called the calculus of variations.

The two sets of ideas both have important applications, but to different problems. At one extreme we have those problems such as the isoperimetric problem of Pappus just mentioned and, more recently, problems in which a function is to be found that produces a best possible result in some sense, such as propelling an airplane between given points with least expenditure of fuel. At the other extreme we have situations in which the presence or absence of a maximum or minimum is irrelevant; only the consequences of stationarity matter. These consequences often include the satisfaction of a set of differential equations. According to what is misnamed "the principle of least action," the motion of a set of particles follows a time-development for which a certain integral, called the "action," is stationary. The function for which the action

* Reprinted with permission (with minor editorial changes) from *Optimal Control and Differential Equations*, A. B. Schwarzkopf, Walter G. Kellet, and Stanley B. Eliason, eds., Academic Press, New York, 1978, pp. 3-51.

† Professor E. J. McShane passed away on June 1, 1989. At the time of his death, he was affiliated with the Department of Mathematics, University of Virginia, Charlottesville, Virginia 22903.

is stationary is the one for which the classical equations of motion are satisfied, and the satisfaction of those equations is all that we want.

In between these two extremes we have the problems of relative extrema. Let us say that a function y is in the weak ε-neighborhood of another function y_0 if there is a homeomorphism between their graphs such that at corresponding points, the values of y and y_0 differ by less than ε, and so do the values of their derivatives. The function y is in the strong ε-neighborhood of y_0 if this holds with the reference to the derivatives deleted. A functional has a weak (strong) relative minimum at y_0 if for some positive ε, the functional has at y_0 its least value on the set of all those y in the domain of the functional that are in the weak (strong) ε-neighborhood of y_0. These concepts have some applications, related to stable and unstable equilibrium; but I have a strong suspicion that relative maxima and minima were usually studied, not because they were really wanted, but because available theory did not permit the study of absolute maxima and minima.

For lack of time I shall say little about the second set of ideas, based on stationarity. This means that I shall disregard some important pure mathematics and some important applications. I have mentioned that the principle of least action is of this type. So too is Hamilton's study of optics and its extension into calculus of variations by Jacobi. So is all the mathematics of quantum theory that is based on a Hamiltonian. So, too, is Marston Morse's theory of the calculus of variations in the large. I shall choose for my principal subject the development of the first set of ideas that I have called the theory of extrema.

In the eighteenth century the distinction between the two sets of ideas was hardly noticed. If it could be shown that any curve that minimized some functional had to satisfy a certain condition, and a curve could be found that did satisfy that condition, it was accepted without comment that that curve did furnish the minimum. Nor has such a feeling quite disappeared. On page 16 of the book by Gelfand and Fomin [1] (English translation) we read: "In fact, the existence of an extremum is often clear from the physical or geometric meaning of the problem, e.g., in the brachistochrone problem, the problem concerning the shortest distance between two points, etc. If in such a case there exists only one extremal satisfying the boundary conditions of the problem, this extremal must perforce be the curve for which the extremum is achieved." I disagree with this on three counts. First, if the calculus of variations is mathematics, our conclusions must be deducible logically from the hypotheses, with no use of anything that is "clear from the physical meaning"—even if anything is ever that clear in physics. Second, if the mathematical expression is meant to be a model of a physical situation, we are not entitled to unshakable confidence that the model we have chosen is perfect in all details; rather, we should keep in mind that a mathematical model of a physical system is necessarily a simplification and idealization. Third, the principle as stated is untrustworthy. For example, if A and B are two points in the upper half-plane, there always exists a curve joining them such that the surface of revolution obtained by rotating it about the x-axis has least area. If A and B are properly located, there is just one extremal that joins them, and it does not furnish the least area. (See G. A. Bliss [2, p. 116].)

In the early eighteenth century the necessary conditions for a minimum in various specific problems were found by ingenious devices, usually involving replacing a short arc of the curve by another short arc with the same ends. In 1760, Lagrange unified these special solutions by means of the idea of a variation. Suppose that a function $x \to y_0(x)$ $(x_0 \leqq x \leqq x_1)$ minimizes a functional $J(y(\cdot))$ in a certain class K of functions. Suppose further that we can find a family of functions $y_\alpha(-b < \alpha < b)$ such that for

each α in $(-b, b)$, the function $x \to y_\alpha(x)(x_{0,\alpha} < x < x_{1,\alpha})$ is in the given class K. Then the derivative at $\alpha = 0$ of the function $J(y_\alpha(\cdot))$, if it exists, must be zero. The function

$$x \to \eta(x) = \partial y_\alpha(x)/\partial\alpha \qquad (\alpha = 0)$$

is often called a variation of y_0; Lagrange used the term "variation" and the symbol δy for the product of this by $d\alpha$. The variation of the functional, which is the derivative of $J(y_\alpha(\cdot))$ at $\alpha = 0$, is the directional derivative of J in the direction η. In many interesting cases its vanishing is equivalent to the satisfaction of a certain differential equation; this is the Euler-Lagrange equation.

For the purposes of mechanics, the goal had now been reached. The Euler-Lagrange equation permitted the introduction of general coordinate systems, and the concept of stationary curve unified the whole theory of classical mechanics, as Lagrange showed in his masterful work. But it was a mental confusion, consistent with the somewhat uncritical ideas of the period, to think that any stationary curve would certainly furnish a maximum or a minimum, as wished. In his *Principia* (1687), Isaac Newton had discussed the problem of finding a surface of revolution with assigned base and altitude that minimized a functional that Newton thought represented the drag when the body is moved through a fluid. Legendre published his necessary condition for a minimum in 1786, a century later; but in 1788, he published another paper, entitled "Mémoire sur la manière de distinguer les maxima des minima dans le calcul de variations," in which he pointed out that a curve could satisfy the Euler-Lagrange equation for the integral expressing the Newtonian resistance and still not give the surface of least resistance. The most interesting feature of his proof is that he showed that the Weierstrass condition for a minimum was not satisfied—and Weierstrass was not born until twenty-seven years later. This work must not have had the immediate effect that it deserved. Mathematicians continued to act as though the only feature of importance was the satisfaction of the condition for stationarity. More than two decades later Robert Woodhouse, F.R.S., a Fellow of Caius College, Cambridge, published a book entitled *Treatise on Isoperimetrical Problems and the Calculus of Variations* (1810), in which Legendre is not mentioned. In this book, Woodhouse poses the problem of maximizing the integral

$$\int [d^2y/dx^2]^2\, dx,$$

the class of curves not being clearly specified. By use of variations, he came to the conclusion that the maximum is provided by the line segment joining the endpoints. Had he used Legendre's results, he would have recognized the falsity of his conclusion. But even without having read Legendre, he should have noticed that unless the endpoints coincide, no maximum can exist, and the line segment gives to the integral the value zero, an obvious minimum.

The guiding principle during the eighteenth century and more than half the nineteenth seemed to be that if a minimizing function is sought for some functional, then by inventing more and more necessary conditions for a minimum, we can feel steadily more confident that a function that passes all the tests is in fact the minimizing function sought. The first necessary condition was stationarity, established when the curve being tested can be varied in arbitrary directions. The next in order of time was the Legendre condition, still in the domain of Lagrange-type variations, and in fact needing only variations that leave the function unchanged outside a small interval. Next came the condition of Jacobi. Like that of Legendre, it expressed the fact that

for a minimum, all directional second derivatives (second variations) must be nonnegative; but unlike Legendre's, it required the variation of the function along long intervals. Next came the necessary condition of Weierstrass. Unlike the others, it cannot be established by means of Lagrange-type variations or directional derivatives. The function being tested is compared with other functions near it in position but widely different in derivative. That is to say, the Weierstrass condition is necessary for a strong relative minimum, not for a weak one.

But Weierstrass made a more significant contribution than the discovery of a new necessary condition. For unconditioned problems, in which the minimum of an integral

$$\int_{x_0}^{x_1} f(x, y(x), y'(x))\, dx$$

is sought in the class of all sufficiently well-behaved functions with assigned end-values, he was able to prove that when a function $y(\cdot)$ satisfies conditions that are slight strengthenings of the four known necessary conditions, it will provide a strong relative minimum for the integral. Now, at last, instead of feeling confident without conclusive proof that a curve gave a minimum to the integral, we could feel certain that it gave a kind of minimum—not indeed the absolute minimum that we were seeking, but at least a strong relative minimum.

This was truly a great step forward in the theory of the calculus of variations. (It might help to promote humility among us workers in that field if we notice that in his biography of Weierstrass in *Men of Mathematics*, E. T. Bell [3] does not even mention that Weierstrass wrote on the calculus of variations.) But it had a psychological drawback. Like the ideas introduced by Lagrange a century earlier, the means used by Weierstrass were so highly esteemed that they became ends in themselves. No matter what the calculus of variations was formally stated to be, in the hands of many of its workers it became a procedure of proving in each new type of problem some analogues of the necessary conditions of Euler and Lagrange, of Legendre, of Jacobi, and of Weierstrass, and then of proving a sufficiency theorem of some sort. This is not astonishing. When I studied calculus, a mere fifty-five years ago, the theory of maxima and minima consisted of finding points at which the derivative of a function is zero and then looking at the value of the second derivative. I learned a needed lesson years later, when for quite practical reasons I needed to find the absolute minimum of a function, and discovered that setting the derivative equal to zero located the maximum; the minimum that I needed was at an endpoint, where the derivative was not zero.

Beginners in calculus today are taught a better method of finding minima of functions f on a closed interval $[a, b]$. First it is shown (or at least asserted in an authoritative tone of voice) that a minimum exists. Next, conditions are found that must be satisfied at the point x_0 at which f is minimum; either x_0 is a or b, or the derivative exists at x_0 and is zero, or the derivative does not exist at x_0. In many problems, these necessary conditions rule out all but a few values of x. One of these gives f its least value; which one can be determined by calculating f at these points. A similar method could be used to find the absolute minimum of a functional provided that first an existence theorem is proved, and then necessary conditions are found that have to be satisfied at the minimizing function. If these necessary conditions rule out all but a few functions, calculating the corresponding values of the functional will permit us to find which one furnishes the absolute minimum.

Early in this century David Hilbert proved an existence theorem for certain unconditioned problems. Later, Leonida Tonelli showed that if an integrand $f(x, y, y')$ is convex as a function of y', its integral is lower semicontinuous; if a sequence of

functions $y_1, y_2, \cdots$ tends in the strong topology to a limit function y_0, the limit inferior of the integral along y_n is at least equal to the integral along y_0. This, with some other very reasonable hypotheses, gave excellent existence theorems for unconditioned problems. But when applied to conditioned problems, such as isoperimetric problems and Bolza problems, it produced no results of interest. These conditioned problems are neither new nor artificial. As to newness, Forsythe asserts that the word "isoperimetric" was first used in the early fifth century by Bishop Synesius. As to artificiality, the engineering problems that led the Russian mathematicians to devise the modern form of control theory are almost invariably conditioned problems. Conditioned problems were unmanageable until L. C. Young invented what he called generalized curves.

For those of us who have not encountered generalized curves, a bit of explanation might be helpful. Suppose that we wish to minimize the integral

$$\int_0^1 [y^2+(y'^2-1)^2]\,dx \tag{1}$$

in the class of absolutely continuous functions $y(\cdot)$ on $[0, 1]$. If we divide $[0, 1]$ into $2n$ intervals of equal length and define y_n to be the function with $y_n(0)=0$ and $y'=1$ and $y'=-1$ on alternate subintervals, the graph of y_n is a sawtooth polygon, and the integral has value $1/12n^2$. So the lower bound of the integral is zero. But y_n tends uniformly to the zero function, for which the integral (1) has value 1. We need a different approach. Let us plot each value of y_n' on a u-axis. If we select a subinterval of $[0, 1]$ and choose an x at random in it, there is a certain probability that $y_n'(x)=1$, and this probability tends to $\frac{1}{2}$ as n increases; and likewise, the probability that $y_n'(x)=-1$ tends to $\frac{1}{2}$ as n increases. So we construct a new kind of object. Instead of having a number $y'(x)$ associated with each x in $[0, 1]$, it has a probability distribution P_x that assigns probability $\frac{1}{2}$ to each of the numbers 1, -1 and the probability zero to the rest of the real number system. This is what we would have in the unrealizable situation that on every subinterval of $[0, 1]$, $y'(x)$ were $+1$ half the time and -1 half the time. This distribution takes the place of the single number $y'(x)$, which can be thought of as a distribution in which probability 1 is assigned to the number $y'(x)$ and zero to the rest of the real numbers. Thus instead of having $y_0(x_1)$ equal to the integral of $y_0'(x)$ from zero to $x_1 (0 \leqq x_1 \leqq 1)$, we have

$$y_0(x_1)=\int_0^{x_1}\left\{\int_R uP_x(du)\right\}dx=0;$$

and likewise, the replacement for the integral (1) is

$$\int_0^1\left\{\int_R [y_0^2+(u^2-1)^2]P_x(du)\right\}dx,$$

which has the value zero. Young's generalized curves are objects of this new kind. A generalized curve can be thought of as a pair $((y(x), P_x): a \leqq x \leqq b)$ in which the y is a function on $[a, b]$, and for each x in $[a, b]$, P_x is a probability distribution on R, and

$$y(x_1)=\int_a^{x_1}\left\{\int_R uP_x(du)\right\}dx, \qquad (a \leqq x_1 \leqq b).$$

(This is close to Young's original formulation; in his book *Calculus of Variations and Optimal Control Theory* [4], he prefers to regard a generalized curve as a functional on a class of integrands, somewhat like Schwartz distributions.) This can be generalized at once to higher dimensions.

The space of generalized curves can be topologized by defining the statement that a sequence of generalized curves $((y_n(x), P_{n,x})\colon a_n \leqq x \leqq b_n)$ tends to a generalized curve $((y_0(x), P_{0,x})\colon a_0 \leqq x \leqq b_0)$ if and only if

$$\lim_{n\to\infty} \int_{a_n}^{b_n} \left\{ \int_R \phi(x, y_n(x), u) P_{n,x}(du) \right\} dx = \int_{a_0}^{b_0} \left\{ \int_R \phi(x, y_0(x), u) P_{0,x}(du) \right\} dx$$

for every continuous function ϕ that vanishes outside a bounded set. The extension to higher dimensions is obvious. The remarkable fact is that with this topology, the space of generalized curves has sufficiently strong compactness properties so that for a large class of problems, not merely unconditioned problems, a minimizing generalized curve can be found for the integral under consideration. Young applied this in 1937 to unconditioned problems. In 1940, I published a sequence of three papers in which, for problems of Bolza in parametric form, it was shown first, that under weak hypotheses a minimizing generalized curve exists; second, that it satisfies necessary conditions that are generalizations of the Euler-Lagrange, Legendre and Weierstrass conditions; and third, that under some extra hypotheses, the minimizing generalized curve has each probability measure P_x concentrated at a single point, so that it is in fact an ordinary curve in another notation [5]-[7].

This set of papers burst on the mathematical world with all the éclat of a butterfly's hiccough. The reaction of mathematicians was like that of the little boy who wrote his grandmother: "Thank you for the book about penguins. It taught me more than I wanted to know about penguins." Because it provided a means of finding extrema analogous to today's method of treating minima in calculus, it extended to mathematicians the privilege of forgetting about semicontinuity and about sufficiency theorems. But it was superfluous. Without it, almost all of them had already forgotten about semicontinuity and about sufficiency theorems. And they were justified. The problem of Bolza was the most general of the single-integral problems of the calculus of variations. Its mastery gave us the power to answer many deep and complicated questions that no one was asking. The whole subject was introverted. We who were working in it were striving to advance the theory of the calculus of variations as an end in itself, without attention to its relation with other fields of activity.

In contrast, the theory of optimal control attracted great attention as soon as Pontryagin and his followers published it in the late 1950s; and I think that that is as it should be. In my mind, the greatest difference between the Russian approach and ours was in mental attitude. Pontryagin and his students encountered some problems in engineering and in economics that urgently asked for answers. They answered the questions, and in the process they incidentally introduced new and important ideas into the calculus of variations. I think it is excusable that none of us in this room found answers in the 1930s for questions that were not asked until the 1950s. But I for one regret that when the questions arose, I did not notice them. Like most mathematicians in the United States, I was not paying attention to the problems of engineers.

In order to discuss this new aspect of the calculus of variations, it is convenient to introduce a different method of formulating extremum problems that includes all the older formulations and also the new problems that arose in the 1950s. A curve $t \to y(t)$ can be regarded as the path of a moving point, and the motion can be controlled by choosing a value of $\dot{y}(t)$ for each t. But with conditioned problems, we may not be able to choose $\dot{y}$ arbitrarily. For example, the n components of the vector $\dot{y}(t)$ may have to satisfy some differential equations, fewer than n of them. Also, we may not wish to choose $\dot{y}(t)$ directly, but to fix it by choosing some parameters u that determine

$\dot{y}$. Therefore, we shall suppose that there exist $n+1$ functions $(t, y, u) \to f^i(t, y, u)$, defined for all (t, y) in $(n+1)$-space R^{n+1} and all u in a set $\Omega(t, y)$. We control the curve by choosing a function $t \to u(t)$. If there is an n-vector-valued function $t \to y(t)$ $(a \leqq t \leqq b)$ such that for all t in $[a, b]$, $u(t) \in \Omega(t, y(t))$ and

$$y^i(t) = y^i(a) + \int_a^t f^i(\tau, y(\tau), u(\tau))\, d\tau, \qquad (i = 1, \cdots, n), \tag{2}$$

the function u is an admissible control function and $y(\cdot)$ is the response, or trajectory, corresponding to it. The problem is to find an admissible control $u(\cdot)$ such that the integral

$$\int_a^b f^0(t, y(t), u(t))\, dt \tag{3}$$

attains its least value among all admissible controls for which the responses satisfy certain end-conditions.

This formulation not only includes all previous ones; it is easier to work with. Magnus Hestenes stated it in 1950; but his results were published in a RAND report with little circulation and at a time when the calculus of variations was at ebb-tide, and they did not attract the attention they deserved. Even earlier, in a paper published in the *Transactions of the American Mathematical Society* in 1933, L. M. Graves had transformed the problem of Lagrange into the control formulation and had established analogues of the Lagrange multiplier rule (the Euler-Lagrange equation) and the Weierstrass condition. These together are equivalent to the "Pontryagin maximum principle," but only when the set $\Omega(t, y)$ is all of a Euclidean space and the minimizing curve satisfies an annoying condition called "normality." Apparently Pontryagin and his associates did not notice that Graves and Hestenes had both made use of the notation, and they invented it independently. But they introduced one important new feature. They allowed the sets $\Omega(t, y)$ to be closed sets, not demanding that they be open as previous researchers had. To me, as a participant in the older research, it is of interest to distinguish what it was that they adapted from work of their predecessors and what they introduced that was quite new. But I lack the time, and I suspect that most of us lack the interest, for such historical research. However, I do wish to point out that in their book, published in English translation in 1962, Pontryagin and his co-authors made a great step forward in one respect, but a step backward in another. They stated a "maximum principle" that is a generalization of the necessary conditions of Euler and Lagrange, of Legendre and of Weierstrass; but like the mathematicians of the eighteenth century, they gave no sufficient conditions for a minimum, and they stated an existence theorem only for a quite special and simple case. Except for this last theorem, their theory is what L. C. Young calls the "naive" theory.

Other authors proved existence theorems, but usually under the restrictive condition that the image in R^{n+1} of $\Omega(t, y)$ under the mapping

$$u \to (f^0(t, y, u), \cdots, f^n(t, y, u))$$

is a convex set. Far better results can be obtained by generalizing the problem to allow "relaxed" controls, an obvious extension of the idea of generalized curves. For each t, instead of choosing a point $u(t)$ in $\Omega(t, y(t))$, we choose a probability measure P_t on $\Omega(t, y(t))$. Then equations (2) are replaced by

$$y^i(t) = y^i(a) + \int_a^t \left\{ \int_{\Omega(\tau, y(\tau))} f^i(\tau, y(\tau), u) P_\tau(du) \right\} d\tau \tag{4}$$

and the integral to be minimized is

$$\int_a^b \left\{ \int_{\Omega(\tau, y(\tau))} f^0(\tau, y(\tau), u) P_\tau(du) \right\} d\tau. \tag{5}$$

Since we shall mention these often, it is expedient to introduce some notation and terminology. We shall sometimes write $P_\cdot$ for the distribution-valued function $t \to P_t$. An *admissible pair* is a pair

$$C = (y(\cdot), P_\cdot) = ((y(t), P_t): a \leqq t \leqq b)$$

in which for each t in $[a, b]$, P_t is a probability distribution on $\Omega(t, y(t))$, and the functions $(f^i(t, y(t), u): u \in \Omega(t, y(t))$ are defined and are integrable with respect to P_t over $\Omega(t, y(t))$, and equations (4) are satisfied. The integral (5) will be denoted by $J(C)$:

$$J(C) = \int_a^b \left\{ \int_{\Omega(t, y(t))} f^0(t, y(t), u) P_t(du) \right\} dt.$$

For simplicity we shall restrict our attention to the important special case in which $\Omega(t, y)$ is independent of t and y; we denote it simply by Ω. We shall suppose that F^* denotes a closed set in R^{n+1} and E^* a bounded closed set in R^{2n+2}; the problem is to find an admissible pair C such that the points $(t, y(t))$ lie in F^* and the endpoints $(a, y(a), b, y(b))$ in E^*, and $J(C)$ is the least value of J for all such admissible pairs. If Ω is compact and the f^i are continuous, and some condition is satisfied that guarantees the existence of a minimizing sequence that lies in a bounded subset of F^*, the minimum can be shown to exist. The proof can be found in several research papers and several books in varying degrees of generality and simplicity.

However, in order to carry out our suggested program of solving the minimizing problem, the necessary conditions for an optimum relaxed control should apply to the kind of optimum that has been shown to exist. This is not taken care of in all books and papers on the subject. However, it was done in 1962 in three independent papers, by J. Warga [8], by T. Wazewski [9], and by R. V. Gamkrelidze [10]. It is easily accessible in the books on control theory by L. C. Young [4] and by J. Warga [11]. Young proves an existence theorem somewhat more general than that of the preceding paragraph. For brevity, we shall restrict its generality, but allow its extension to problems in parametric form. A control problem is in parametric form if the control set Ω is a cone, so that whenever u is in Ω so is pu for all nonnegative p; and the end-conditions are independent of t; and the functions f^i are independent of t and are positively homogeneous of degree 1 in u, so that

$$f^i(y, pu) = pf^i(y, u)$$

for u in Ω and $p \geqq 0$. It can then be shown that if F^*, E^*, and Ω are closed and E^* is bounded, and Ω is independent of y, and the f^i are continuous on $F^* \times \Omega$, and either Ω is bounded or the problem is in parametric form, and there exists a minimizing sequence of admissible pairs for which

$$\int_a^b \left\{ \int_\Omega |u| P_t(du) \right\} dt$$

is bounded, then there exists an admissible pair $C = (y(\cdot), P_\cdot)$ for which the integral $J(C)$ has its least value. Moreover, if the problem is in parametric form, the range of t can be taken to be $[0, 1]$, and the probability distribution P_t can be so chosen that for a certain positive number L the support of P_t is in $\{u \in \Omega: |u| = L\}$; that is,

$$P_t\{u \in \Omega: |u| \neq L\} = 0 \qquad (0 \leqq t \leqq 1).$$

For discussing necessary conditions for optimal controls, we consider only the case in which for the optimizing pair $(y(\cdot), P_\cdot)$ the points $(t, y(t))$ lie in the interior of the set F^*. Also, for simplicity we shall consider only the fixed-end-point problem. For these it is proven, in Young's book and elsewhere, that the following theorem holds.

THEOREM. *Let* $((y(t), P_t): a \leqq t \leqq b)$ *be an optimal pair for the problem described above. For each* t *in* $[a, b]$, *let* $(t, y(t))$ *be interior to* F^*, *and the support of* P_t *in a bounded subset of* Ω. *Then there exist a constant* ψ_0, *with value zero or* -1, *and* n *absolutely continuous functions* $\psi_1, \cdots, \psi_n$ *on* $[a, b]$, *with the following properties.*

(i) $\psi_0, \psi_1(t), \cdots, \psi_n(t)$ *are not all zero for any* t *in* $[a, b]$.

(ii) *If for each* t *in* $[a, b]$, *each* y *in* F^* *and each* u *in* Ω *we define*

$$H(t, y, u) = \psi_0 f^0(t, y, u) + \sum_{i=1}^{n} \psi_i(t) f^i(t, y, u),$$

then the maximum value of $H(t, y(t), u)$ *on* Ω *is a continuous function of* t *on* $[a, b]$; *and if the problem is in parametric form, the value of this maximum is zero.*

(iii) *For almost all* t *in* $[a, b]$, *the support of* P_t *is contained in the set on which* $H(t, y(t), u)$ *attains its maximum value*; *and by changing* P_t *on at most a set of measure zero, we can cause this to hold for all* t *in* $[a, b]$.

(iv) *The functions* ψ_i *satisfy*

$$d\psi_i(t)/dt = -\int_\Omega \frac{\partial H}{\partial y^i}(t, y(t), u) P_t(du)$$

for almost all t *in* $[a, b]$.

This theorem and its generalizations are well known to have interesting applications in problems of optimal control. Many of them are contained in the books by Pontryagin et al., by L. C. Young, and by L. D. Berkovitz [12]. But if the optimal control formulation is, as I believe, the modern replacement for the classical calculus of variations, it must be able to provide solutions for the problems of the classical calculus of variations, and it should do so with at most little additional effort. Doubts have been expressed that the ancient problems can be at all conveniently solved by optimal control methods, without transferring back to the old notation. I do not think that that is so, and to bear out my opinion I have worked out several very old problems by optimal control theory. Since the point of the task is to show that a treatment in full detail can be presented without especial difficulty, I lack time and space to present all these calculations in this talk. Instead, I have prepared some sheets [the Appendix] on which I have written out the details of the solutions of four classical problems. Although these problems have been discussed in many books, the treatment there is "naïve," with no existence theorems being established. More surprisingly, even from the naïve point of view many of the discussions are incomplete or incorrect. The fourth problem is to me the most interesting. It is the ancient problem of Newton: to find a surface of revolution for which a certain integral, believed by Newton to represent the resistance encountered by that surface when moved through a fluid, has least possible value. Newton did not clearly specify the curves he permitted. So in practically all discussions, all piecewise smooth functions $x \to y(x)$ are allowed, and it is then shown that the problem is unsolvable. Only Goursat, in the third volume of his *Cours d'Analyse*, points out that it is physically reasonable to assume y monotonic; this might well have been what Newton meant. In no book, not even in Goursat's, is Newton's statement about the solution quoted in full. But the solution of the problem with monotone y exists, and it can be found by the methods of optimal control theory, and it is just

what Newton said it is. In this problem it appears that the optimal control procedures are essential; the solution is not accessible by classical procedures.

From what I have just said, it is easy to deduce what I believe should be done about the teaching of the calculus of variations. Ordinary undergraduate students of mathematics should be taught a form of control theory simple enough to be understood and general enough to be applicable to many problems. This may call for some quite new chapters in advanced calculus texts, but I think it is not unattainable, and not even very difficult. But another quite different matter is the direction of research. During the twentieth century the interest in what might be called traditional calculus of variations has sunk to a low ebb. I think that that is a natural consequence of the introversion of the subject. Theorems were proved of increasing intricacy, of interest to a steadily shrinking collection of experts in the subject. The newer calculus of variations will go the same way unless its practitioners are sensitive to the questions that arise naturally and demand answers. My own guess is that some of these have to do with the consideration of problems in which random events play an important part. There has, in fact, been a considerable development of stochastic control theory. But both in it and in the deterministic theory I feel that the tendency toward introversion is showing up. The theorems are becoming more baroque. In particular, in quite complicated situations we can show that a solution exists; but only in simple situations can we find a usable approximation to that solution without vast computational effort. The situation is bad in the deterministic case and worse in the stochastic. I am no expert in computation, but I have been told that the direct application of the maximum principle to problems of even moderate complexity is unsatisfactory. This is easy to believe, because the direct application of the maximum principle would require us to find the value of u at which $H(t, y(t), u)$ takes its maximum value, and it is hard to find precisely where a maximum occurs. For stochastic problems, the situation is worse. In some cases solutions have been stated which in my opinion are not solutions at all, since they ask the controller to perform infinitely many adjustments of the controls guided by the instantaneous availability of infinitely many bits of information. Others, for example, Balakrishnan, have worked on the problem of finding approximate solutions that are humanly attainable. But a great deal remains to be done both in the deterministic and in the stochastic cases. A friend of mine, a logician, once gave a talk in which he proved that no matter how many problems in mathematics have been solved at any given time, there will always remain unanswered questions. We hardly need that theorem. Just within optimal control theory there are *good* problems enough to fill all the time and demand all the brain power that all the available mathematicians can give to them.

Appendix: Four classical problems of the calculus of variations

The four problems we shall solve are in parametric form. In each, F^* is a closed set in n-space, and y_0 and y_1 are points of n-space, and Ω is a closed cone in r-space. The functions

$$(y, u) \to f^i(y, u) \qquad (i = 1, \cdots, n)$$

are continuous on $F^* \times \Omega$ and are positively homogeneous of degree 1 in u for each fixed y. The minimum of

$$J(C) = \int_a^b \left\{ \int_\Omega f^0(y(t), u) P_t(du) \right\} dt$$

is sought in the class of all admissible pairs $((y(t), P_t)\colon a \leqq t \leqq b)$ for which every $y(t)$

is in F^*, and

$$y^i(t) = y_0^i + \int_a^t \left\{ \int_\Omega f^i(y(\tau), u) P_\tau(du) \right\} d\tau \qquad (a \leqq t \leqq b),$$

and

$$y^i(b) = y_1^i.$$

For such problems it is known that a minimizing pair $((y(t), P_t): 0 \leqq t \leqq 1)$ exists, the support of P_t being contained in a set $\{u \in \Omega: |u| = L\}$, provided that there exists a minimizing sequence of admissible pairs $((y_n(t), P_{n,t}): a_n \leqq t \leqq b_n)$ for which the integrals

$$\int_{a_n}^{b_n} \left\{ \int_\Omega |u| P_{n,t}(du) \right\} dt$$

are bounded.

It can also be shown, as a corollary of the "maximum principle," that if $((y(t), P_t): a \leqq t \leqq b)$ is a minimizing pair for this problem, and for each t in $[a, b]$ the support of P_t is bounded and $y(t)$ is interior to F^*, there exist a constant ψ_0 $(=0$ or $-1)$ and n absolutely continuous functions on $[a, b]$, called $\psi_1, \cdots, \psi_n$, such that $\psi_0, \psi_1(t), \cdots, \psi_n(t)$ never vanish simultaneously and

(i) for t in $[a, b]$ the maximum value of

$$H(t, y, u) = \psi_0 f^0(y, u) + \sum_{i=1}^n \psi_i(t) f^i(y, u) \qquad (u \in \Omega)$$

is zero;

(ii) for almost all t in $[a, b]$, this maximum is attained at each point of the support of P_t;

(iii) $$\psi_i(t) = \psi_i(a) - \int_a^t \left\{ \int_\Omega \frac{\partial H}{\partial y^i}(\tau, y, u) P_\tau(du) \right\} d\tau$$

$(a \leqq t \leqq b)$.

In all four problems $r = 2$, and to avoid superscripts, we shall write (u, v) instead of (u^1, u^2). Likewise, points in two-space shall be denoted by (x, y) instead of (y^1, y^2), and points in three-space by (x, y, z).

1. The classical isoperimetric problem. Given a line segment AB, we are to find a curve of length L_0 that goes from B to A and together with AB encloses the greatest possible area. (See Fig. 1.)

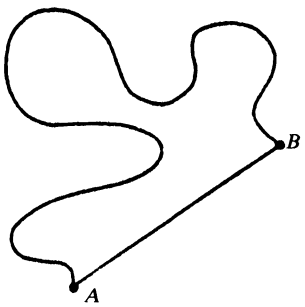

FIG. 1

If the curve is $x = x(t)$, $y = y(t)$, $a \leqq t \leqq b$, the area enclosed is constant $+ \int_a^b [x\dot{y} - y\dot{x}]\, dt$.

We restate the problem in relaxed-control form. We denote B, A by (x_0, y_0), (x_1, y_1), respectively; we define Ω to be the (u, v)-plane; and we define

$$f^0(x, y, z, u, v) = yu - xv,$$

$$f^1(x, y, z, u, v) = u,$$

$$f^2(x, y, z, u, v) = v,$$

$$f^3(x, y, z, u, v) = |(u, v)| = [u^2 + v^2]^{1/2}.$$

The problem is to find an admissible pair $C = ((x(t), y(t), z(t), P_t)\colon a \leqq t \leqq b)$ with

$$\begin{aligned} &x(a) = x_0, \quad y(a) = y_0, \quad z(a) = 0, \\ &x(b) = x_1, \quad y(b) = y_1, \quad z(b) = L_0, \end{aligned} \tag{6}$$

for which the integral

$$J(C) = \int_a^b \left\{ \int_\Omega f^0(x(t), y(t), z(t), u, v) P_t(du, dv) \right\} dt$$

is minimum. Since the endpoints are fixed and the lengths bounded, a solution exists, by the existence theorem on page 926. For it we can assume that for a certain L, the support of P_t is on the circle $|(u, v)| = L$. Since F^* is the whole (x, y, z)-space, the necessary conditions on page 926 hold; H has the form

$$H(t, x, y, z, u, v) = \psi_0[yu - xv] + \psi_1(t)u + \psi_2(t)v + \psi_3(t)|(u, v)|.$$

In particular, H is independent of z, so by (iii) ψ_3 is a constant. If this constant were zero, we would have

$$H(t, x, y, z, u, v) = [\psi_0 y + \psi_1(t)]u + [-\psi_0 x + \psi_2(t)]v.$$

In order for this linear function to have a maximum on Ω the coefficients of u and v must be zero, so that

$$\psi_1(t) = -\psi_0 y, \qquad \psi_2(t) = \psi_0 x.$$

If $\psi_0 = 0$, all four ψ_i are zero, which is false. If $\psi_0 = -1$, these last equations are incompatible with (iii). So $\psi_3 \neq 0$.

We consider two cases.

Case 1. $\psi_0 = 0$.

By (iii), ψ_2 and ψ_1 are both constants. Then H assumes its maximum on the circle $|(u, v)| = L$ at just one point, independent of t, so $\dot{x}$ and $\dot{y}$ are constants and $(x(t), y(t))$ traverses a line segment from B to A. This is possible, with the end-conditions (6), if and only if L_0 is equal to the distance from B to A. In this case, we have shown that the line segment BA gives the maximum area. But this is obvious without the discussion, since then there is only one curve that satisfies the conditions for admissibility.

Case 2. $\psi_0 = -1$.

In this case,

$$H(t, x, y, z, u, v) = [-y + \psi_1(t)]u + [x + \psi_2(t)]v + \psi_3(t)|(u, v)|. \tag{7}$$

For (u, v) on the circle $|(u, v)| = L$, this has maximum value at just one point. So for each t, the support of P_t consists of a single point, and the optimal pair is ordinary. By (iii) on page 926,

$$d\psi_1(t)/dt = -\int_\Omega [\partial H/\partial x] P_t(du, dv) = -\int_\Omega v P_t(du, dv) = -\dot{y}(t)$$

for almost all t. Likewise, for almost all t,

$$d\psi_2(t)/dt = \dot{x}(t).$$

Therefore there exist constants c_1, c_2 such that

$$\psi_1(t) = 2c_2 - y(t), \qquad \psi_2(t) = x(t) - 2c_1.$$

We substitute this in (7). By (iii), when P_t has only one point in its support, that point is $(\dot{x}(t), \dot{y}(t))$, so H has its maximum there, and its partial derivatives as to u and v are zero. This yields

$$2c_2 - 2y(t) + \psi_3 \dot{x}(t)/L = 0,$$
$$2x(t) - 2c_1 + \psi_3 \dot{y}(t)/L = 0.$$

Therefore

$$\frac{d}{dt}[(x(t) - c_1)^2 + (y(t) - c_2)^2] = 2(x(t) - c_1)\dot{x}(t) + 2(y(t) - c_2)\dot{y}(t) = 0,$$

and the optimal curve is a circular arc with endpoints B and A.

It should be observed that there may be several such arcs. For example, in Fig. 2, $BCADEBCA$ is one, and so is BCA. But an arc that passes more than once through A yields the same area as a curve obtained by rotating through 180° a loop beginning and ending at A, as in Fig. 2(b). This new curve $BCAD'E'A$ does not furnish the maximum area, because it is not a circular arc. Therefore, neither did the multiply-traversed circular arc from which we obtained it. The curve that encloses the greatest area is a circular arc without multiple points that goes from B to A.

2. The brachistochrone. We next consider the problem of the brachistochrone, first proposed by John Bernouilli in 1696.

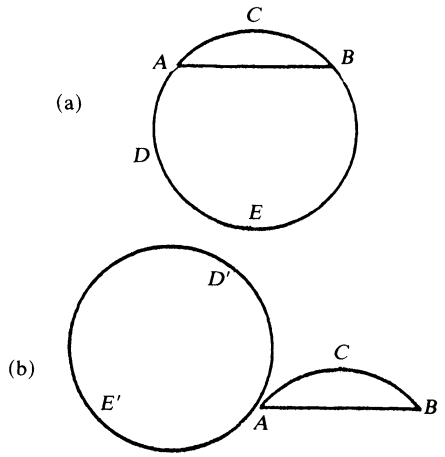

FIG. 2

Let us choose axes in the plane with the y-axis vertically downward. A bead descends by gravity, starting from rest, along a frictionless wire beginning at $(0, 0)$ and ending at a point (x_1, y_1) with $x_1 > 0$ and $y_1 \geqq 0$. It is required to find the shape of the wire along which the time of descent will be shortest.

In order for the wire to be traversible at all it must lie in the half-plane

$$F^* = \{(x, y): y \geqq 0\}.$$

Along a curve $x = x(t)$, $y = y(t)$ $(a \leqq t \leqq b)$ in F^* with absolutely continuous $x(\cdot)$ and $y(\cdot)$ the time of descent is proportional to

$$\int_a^b \left[\frac{\dot{x}(t)^2 + \dot{y}(t)^2}{y(t)}\right]^{1/2} dt;$$

we are to minimize this. We extend the problem to relaxed-control form. We define F^* as above, and define Ω to be the (u, v)-plane. Admissible pairs are those pairs $C = ((x(t), y(t), P_t): a \leqq t \leqq b)$ that satisfy

$$x(t) = \int_a^t \left\{\int_\Omega u P_\tau(du, dv)\right\} d\tau,$$

$$y(t) = \int_a^t \left\{\int_\Omega v P_\tau(du, dv)\right\} d\tau$$

and the end-conditions

$$x(b) = x_1, \qquad y(b) = y_1,$$

and have $(x(t), y(t))$ in F^* for all t. Among these we seek a pair C for which the integral

$$J(C) = \int_a^b \left\{\int_\Omega f^0(x(t), y(t), u, v) P_t(du, dv)\right\} dt$$

is a minimum, where

$$f^0(x, y, u, v) = [u^2 + v^2]^{1/2}/y^{1/2}.$$

Although the existence theorem on page 926 does not apply to this problem, the integrand being discontinuous at $y = 0$, a slight modification using a limit process can be used to show that a minimizing pair exists, with $a = 0$, $b = 1$, and the support of P_t contained in the set $\{|(u, v)| = L\}$. We omit the details of this proof.

Suppose that there is a t^* in the open interval $(0, 1)$ for which $y(t^*) = 0$. Since $J(C)$ is finite, y is not identically zero on $[0, t^*]$ or on $[t^*, 1]$. We can therefore find a positive number c and numbers t_1, t_2 such that $0 < t_1 < t^* < t_2 < 1$, and $y(t_1) = y(t_2) = c$, and $y(t) < c$ on (t_1, t_2). Then

$$\int_{t_1}^{t_2} \left\{\int_\Omega y^{-1/2}[u^2 + v^2]^{1/2} P_t(du, dv)\right\} dt > \int_{t_1}^{t_2} \left\{\int_\Omega c^{-1/2} u P_t(du, dv)\right\} dt$$

$$= \int_{t_1}^{t_2} c^{-1/2} \dot{x}(t)\, dt.$$

So if we replace C by a pair C^* in which on the interval $[t_1, t_2]$ the relaxed pair $(x(\cdot), y(\cdot), P_\cdot)$ is replaced by the ordinary curve $x = x(t)$, $y = c$, C^* is an admissible pair, and $J(C^*) < J(C)$, which is impossible. So $y(t) > 0$ for $0 < t < 1$.

For every subinterval $[a, b]$ of the open interval $(0, 1)$, the admissible pair $((x(t), y(t), P_t)\colon a \leqq t \leqq b)$ minimizes the integral

$$\int_a^b \left\{ \int_\Omega y(t)^{-1/2}[u^2+v^2]^{1/2} P_t(du, dv) \right\} dt$$

in the class of admissible pairs with the same endpoints, and the points of its trajectory lie in the interior of F^*. Therefore, the necessary conditions on page 926 are satisfied. H has the form

$$H(t, x, y, u, v) = \psi_0 y^{-1/2}[u^2+v^2]^{1/2} + \psi_1(t)u + \psi_2(t)v,$$

and it attains its maximum at each point in the support of P_t. If ψ_0 were zero, H could have no maximum, so $\psi_0 = -1$. This implies that H can have its maximum value at only one point of the circle $|(u, v)| = L$, so the curve is ordinary, and the one point in the support of P_t is (by (iii)) the point $(\dot{x}(t), \dot{y}(t))$ for almost all t.

By the first of equations (iii), $\psi_1(t)$ is a constant. If it were zero, the maximum value of H on $|(u, v)| = L$ would occur at $(0, L)$ or at $(0, -L)$, yielding $\dot{x}(t) = 0$ for almost all t. But then $x(b) = x(a)$, and by letting a tend to zero and b to 1 we find $x(1) = x(0)$, which is false. So $\psi_1 \neq 0$.

For almost all t, $(\dot{x}(t), \dot{y}(t))$ is the only point in the support of P_t, so $H(t, x(t), y(t), u, v)$ attains its maximum at that point. Therefore at $(\dot{x}(t), \dot{y}(t))$

$$0 = \partial H/\partial u = -y(t)^{-1/2}\dot{x}(t)[\dot{x}(t)^2+\dot{y}(t)^2]^{-1/2} + \psi_1.$$

Since $\psi_1 \neq 0$, this implies that $\dot{x}(t)$ is bounded away from zero and has the same sign as ψ_1 (necessarily positive). So the function $t \to x(t)$ $(0 \leqq t \leqq 1)$ has an inverse $x \to t(x)$ $(0 \leqq x \leqq x_1)$, and the optimizing curve has the representation

$$x \to Y(x) = y(t(x)) \qquad (0 \leqq x \leqq x_1).$$

By the preceding equation,

$$Y(x)[1+(dY/dx)^2] = \psi_1^{-2}.$$

This is a familiar equation. Its solution is a cycloid; see, for example, El'gol'ts [13, p. 38].

3. The surface of revolution of least area. Euler proposed the problem of finding the curve that joins two points in the (x, y)-plane and, among such curves, generates when revolved about the x-axis that surface that has least area. If these endpoints are (x_0, y_0) and (x_1, y_1), and the curve has a representation $x = x(t)$, $y = y(t)$, $(a \leqq t \leqq b)$ in which $x(\cdot)$ and $y(\cdot)$ are absolutely continuous, the area of the surface of revolution is

$$2\pi \int_a^b |y(t)|[\dot{x}(t)^2+\dot{y}(t)^2]^{1/2}\, dt; \tag{8}$$

and this is the integral to be minimized. It is rather obvious that this is the same problem as minimizing the integral

$$\int_a^b y(t)[\dot{x}(t)^2+\dot{y}(t)^2]^{1/2}\, dt \tag{9}$$

in the class of absolutely continuous functions x, y with the given endpoints and lying in the half-plane

$$F^* = \{(x, y)\colon y \geqq 0\}; \tag{10}$$

see, for example, Bliss' Carus monograph [2, p. 89]. This apparently minor point has led to false assertions in several books. For example, both in Gelfand and Fomin [1] and in El'gol'ts [13] we find the statement that the area is given by (8) with the absolute value sign omitted; this is incorrect, and without the qualification $y \geqq 0$ (which they do not mention) the integral (9) has no lower bound. Bliss correctly states the problem, but on p. 90 says that the necessary conditions previously deduced apply without change to this problem, which is incorrect because the minimizing curve can lie in part along the boundary of F^*.

We restate the problem as a relaxed-control problem. Let F^* be defined by (10), and let Ω be the (u, v)-plane. Define

$$f^0(x, y, u, v) = y[u^2+v^2]^{1/2} \qquad ((x, y) \text{ in } F^*, \text{ all } (u, v)).$$

Among all admissible pairs $((x(t), y(t), P_t)\colon a \leqq t \leqq b)$ that have $(x(t), y(t))$ in F^* for all t, and have endpoints $x(a) = x_0, y(a) = y_0, x(b) = x_1, y(b) = y_1$, and satisfy equations

$$\begin{aligned} x(t) &= x_0 + \int_a^t \left\{ \iint_\Omega uP_\tau(du, dv) \right\} d\tau, \\ y(t) &= y_0 + \int_a^t \left\{ \iint_\Omega vP_\tau(du, dv) \right\} d\tau, \end{aligned} \tag{11}$$

we wish to find one that minimizes the integral

$$J(C) = \int_a^b \left\{ \iint_\Omega f^0(x(t), y(t), u, v) P_t(du, dv) \right\} dt.$$

Let m be the infimum of $J(C)$ in the class of pairs admitted. A minimizing sequence is a sequence of admissible pairs

$$C_n = ((x_n(t), y_n(t), P_{n,t})\colon a_n \leqq t \leqq b_n) \qquad (n = 1, 2, 3, \cdots)$$

satisfying the requirements and such that the integrals $J(C_n)$ tend to m. We distinguish two cases.

Case 1. There exists a minimizing sequence for which the minimum value of $y_n(t)$ on $[a_n, b_n]$ is arbitrarily near zero.

In this case we can choose a subsequence (which without loss of generality we may take to be the original sequence) for which

$$\lim_{n\to\infty} \min \{y_n(t)\colon a_n \leqq t \leqq b_n\} = 0. \tag{12}$$

Let c_n be a point in $[a_n, b_n]$ at which y_n attains its least value. Then

$$\begin{aligned} \int_{a_n}^{c_n} \left\{ \iint_\Omega y_n(t)[u^2+v^2]^{1/2} P_{n,t}(du, dv) \right\} dt &\geqq \int_{a_n}^{c_n} \left\{ \iint_\Omega y_n(t)[-u] P_{n,t}(du, dv) \right\} dt \\ &= -\int_{a_n}^{c_n} y_n(t)\dot{y}_n(t)\, dt = [y_0^2 - y_n(c_n)^2]/2, \end{aligned}$$

and similarly,

$$\begin{aligned} \int_{c_n}^{b_n} \left\{ \iint_\Omega y_n(t)[u^2+v^2]^{1/2} P_{n,t}(du, dv) \right\} dt &\geqq \int_{c_n}^{b_n} \left\{ \iint_\Omega y_n(t) u P_{n,t}(du, dv) \right\} dt \\ &= \int_{c_n}^{b_n} y_n(t)\dot{y}_n(t)\, dt = [y_1^2 - y_n(c_n)^2]/2. \end{aligned}$$

By adding these we obtain

$$[y_0^2+y_1^2-2y_n(c_n)^2]/2 \leqq J(C_n).$$

Since $y_n(c_n)$ is the least value of y_n and $J(C_n)$ tends to m, by (12)

$$[y_0^2+y_1^2]/2 \leqq m.$$

The left member of this inequality is $J(C)$ for the ordinary curve C which is the polygon with successive vertices (x_0, y_0), $(x_0, 0)$, $(x_1, 0)$, (x_1, y_1), so that polygon minimizes $J(C)$ in the class of admissible pairs with the given endpoints.

Case 2. For each minimizing sequence, all y_n have a common positive lower bound.

Choose a minimizing sequence with the same notation as before. Let $c>0$ be a lower bound for all the y_n. Then

$$\int_{a_n}^{b_n}\left\{\int_\Omega [u^2+v^2]^{1/2}P_{n,t}(du, dv)\right\}dt \leqq c^{-1}\int_{a_n}^{b_n}\left\{\int_\Omega y_n(t)[u^2+v^2]^{1/2}P_{n,t}(du, dv)\right\}dt$$

$$= J(C_n)/c.$$

So the integrals are bounded, and by the existence theorem on page 926 a minimizing admissible pair exists, and it can be chosen to be a pair for which $a=0$ and $b=1$, and the support of P_t is contained in a circle $|(u, v)| = L$. For this pair, the minimum of y is positive; otherwise, a sequence of infinitely many repetitions of C would be a minimizing sequence in which the minima of the y_n are arbitrarily near zero, and in the case we are considering that cannot happen. Since $y(t)$ is always positive, the trajectory lies in the interior of F^*, and the necessary conditions on page 926 must be satisfied. Let ψ_0 (=0 or -1), $\psi_1(t)$, $\psi_2(t)$ be the multipliers, so that

$$H(t, x, y, u, v) = \psi_0 y(t)[u^2+v^2]^{1/2}+\psi_1(t)u+\psi_2(t)v.$$

If ψ_0 were 0 this would be a linear function and have no maximum on Ω. Therefore $\psi_0 = -1$.

On the circle $\{|(u, v)| = L\}$ the function $H(t, x(t), y(t), u, v)$ attains its maximum at only one point, and by (11) this is $(\dot{x}(t), \dot{y}(t))$ for almost all t. So the optimal control is ordinary. Since H is independent of x, by the first of equations (iii), ψ_1 is constant. If this constant were zero, the maximum of H would occur at either $(0, L)$ or at $(0, -L)$, and in either case $\dot{x}(t) = 0$. This is impossible, since it would imply $x(1) = x(0)$, which is false. So $\psi_1 \neq 0$. The maximum value of H occurs at $(\dot{x}(t), \dot{y}(t))$, so its partial derivative with respect to u vanishes at that point:

$$0 = -y(t)[\dot{x}^2(t)+\dot{y}^2(t)]^{-1/2}\dot{x}(t)+\psi_1.$$

The quantity in square brackets has value L^2, so $\dot{x}(t)$ is bounded away from zero, and y can be expressed as a function of x. The preceding equation then yields

$$y(t) = \psi_1[1+(dy/dx)^2]^{1/2}.$$

The solution of this is well known to be

$$y = \psi_1 \cosh\ [(x-h)/\psi_1],$$

where h is a constant. This is the equation of a catenary. It is shown in many places that there are at most two such catenaries through (x_0, y_0) and (x_1, y_1). So the absolute minimum of the surface of revolution is given by the polygon or by one of these at most two catenaries.

4. The solid of revolution of least resistance. In his *Principia*, Isaac Newton discussed the resistance encountered by bodies moving through a fluid. Although he never stated a law of resistance specifically, his reasoning in Book II, Proposition XXXIV, Theorem XXVIII (in the discussion of the resistance of a sphere) leads to the following law:

The resisting pressure at any point of the surface *not sheltered from the fluid by some part of the body* is proportional to the square of the component of the velocity along the normal to the surface.

This is Forsythe's formulation, except for the italicized words, which I have added. They are justified by Newton's reasoning and by his ignoring pressure on the sheltered hemisphere. If a solid of revolution is generated by revolving a curve $x = x(t)$, $y = y(t)$ about the y-axis and moves in the direction of the positive y-axis, and x is not monotonic increasing, there will be sheltered arcs such as ABC (see Fig. 3). By Newton's reasoning, there will be no resisting pressure on this part of the solid, and we can replace the arc ABC by the line segment AC without affecting the resistance. So we may, and henceforth shall, assume that $x(\cdot)$ is nondecreasing.

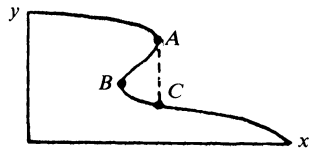

FIG. 3

Newton's assertion about the solid of least resistance is contained in the following two paragraphs of the Scholium following the theorem just cited.

"Incidentally, . . . , it follows from the above that, if the solid $ADBE$ be generated by the convolution of an elliptical or oval figure $ADBE$ about its axis AB, and the generating figure be touched by three right lines FG, GH, HI in the points F, B, and I, so that GH shall be perpendicular to the axis in the point of contact B, and FG, HI may be inclined to GH in the angles FGB, BHI of 135 degrees: the solid arising from the convolution of the figure $ADFGHIE$ about the same axis AB will be less resisted than the former solid, provided that both move forwards in the direction of their axis AB, and that the extremity B of each go forward. This Proposition I conceive may be of use in the building of ships.

"If the figure $DNFG$ be such a curve, that if, from any point thereof, as N, the perpendicular NM be let fall on the axis AB, and from the given point G there be drawn the right line GR parallel to a right line touching the figure in N, and cutting the axis produced in R, MN becomes to GR as GR^3 to $4BR \cdot GB^2$, the solid described by the revolution of this figure about its axis AB, moving in the before-mentioned rare medium from A toward B, will be less resisted than any other circular solid whatsoever, described of the same length and breadth."

Let us equip Newton's figure with an x-axis vertically upward through D and a y-axis along AR, positive in that direction (see Fig. 4). Then B is $(0, y_0)$ and D is $(x_1, 0)$. We denote the length of BG by the perhaps startling symbol $2\psi_2$. The statements in the quoted paragraphs transform into twentieth century notation thus. Let the optimal curve $BGFND$ joining B to D be the graph $x \to y(x)$ $(0 \leqq x \leqq x_1)$. Then $y(x)$ is constantly y_0 for $0 \leqq x \leqq 2\psi_2$; at $2\psi_2$, the right derivative of y is -1; and on $[2\psi_2, x_1]$ the function

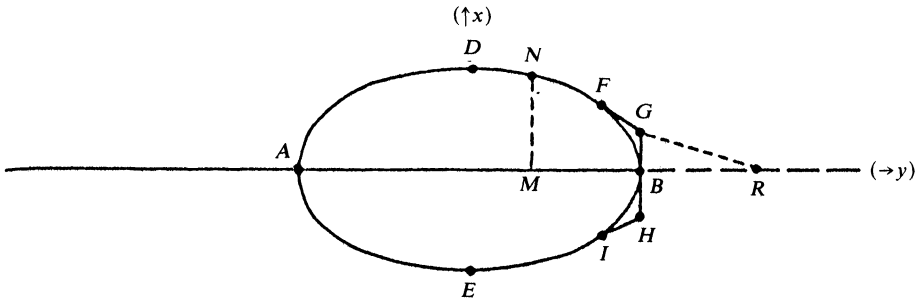

FIG. 4

$y(\cdot)$ satisfies the equation

$$x = -\psi_2[1+(dy/dx)^2]^2/2(dy/dx). \tag{13}$$

In none of the books on the calculus of variations that I have consulted have I found any such statement. They all deduce (13), but omit all reference to the line-segment BG and the corner at G.

Newton did not specify the class of comparison curves allowed. We shall allow only curves $x = x(t)$, $y = y(t)$ in which x is nondecreasing (which, as we have seen, is very nearly implied by Newton's discussion) and $y(\cdot)$ is nonincreasing. Newton might have been willing to accept this reasoning. If y is not monotonic nonincreasing, the solid of revolution will have a trough generated by the revolution of an arc such as DEF (see Fig. 5). When the body moves in the direction of the positive y-axis, this trough will fill with stagnant fluid, and the line-segment DF will be the effective surface, giving a nonincreasing y. Newton makes no such statement. But in the first paragraph quoted above, he specifically considers "oval" solids, and in the Principia and in a letter (presumably written to David Gregory in 1694) only convex figures appear. So we feel that we are not misrepresenting Newton when we attach his name to the following problem:

To find a curve $x = x(t)$, $y = y(t)$, $a \leqq t \leqq b$ in which $x(\cdot)$ is nondecreasing and $y(\cdot)$ is nonincreasing, and

$$x(a) = 0, \quad y(a) = y_0, \quad x(b) = x_1, \quad y(b) = y_1 \qquad (x_1 \text{ and } y_0 \text{ positive}) \tag{14}$$

and which in that class of curves generates the surface of revolution about the y-axis that offers least resistance to motion in the direction of the positive y-axis, the law of resistance being that stated above.

We shall show that this problem has a solution, and that the solution is exactly what Newton asserted it to be.

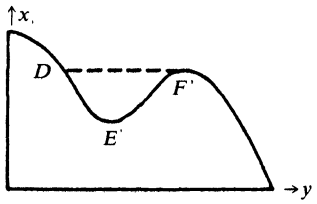

FIG. 5

According to Newton's law of resistance, the resistance at a given velocity is proportional to

$$J(C)=\int_a^b \frac{x(t)\dot{x}(t)^3}{\dot{x}(t)^2+\dot{y}(t)^2}\,dt, \tag{15}$$

and this is the integral to be minimized. We restate the problem in relaxed control form.

Let Ω be the control region

$$\Omega=\{(u, v): u\geqq 0,\ v\geqq 0\}.$$

Define

$$f^0(x, y, u, v)=xu^3/[u^2+v^2],$$
$$f^1(x, y, u, v)=u,$$
$$f^2(x, y, u, v)=-v.$$

A pair $((x(t), y(t), P_t): a\leqq t\leqq b)$ is admissible if P_t is a probability distribution on Ω, and

$$x(t)=\int_a^t\left\{\int_\Omega f^1(x(\tau), y(\tau), u, v)P_\tau(du, dv)\right\}d\tau,$$
$$y(t)=y_0+\int_a^t\left\{\int_\Omega f^2(x(\tau), y(\tau), u, v)P_\tau(du, dv)\right\}d\tau \qquad (a\leqq t\leqq b),$$

and the end-conditions (14) are satisfied. Among all admissible pairs, we seek one that minimizes

$$J(C)=\int_a^b\left\{\int_\Omega f^0(x(t), y(t), u, v)P_t(du, dv)\right\}dt.$$

The simpler existence theorems do not apply to this problem, nor to the analogous problem in ordinary controls, because f^0 is not a convex function. But by the existence theorem on page 926, an admissible (relaxed) pair does exist that minimizes $J(C)$ among all such pairs. It has also the property that $a=0$, and $b=1$, and for all t, the support of P_t is contained in the circular arc

$$\Omega_L=\{(u, v): 0\leqq u, 0\leqq v, u^2+v^2=L^2\}.$$

Since F^* in this problem is the whole plane, the necessary conditions on page 926 are satisfied. As before, we define

$$H(t, x, y, u, v)=\psi_0 f^0(x, y, u, v)+\sum_{i=1}^{2}\psi_i(t)f^i(x, y, u, v);$$

the ψ_0 and ψ_i have the properties on page 926. In particular, since H is independent of y, ψ_2 is a constant.

We shall first prove that $\psi_0\neq 0$. Suppose it were zero. Then

$$H=\psi_1(t)u-\psi_2 v.$$

If H were positive at some (u, v) in Ω, it would be unbounded on the set of points (pu, pv) with $p>0$, which is impossible because H has a maximum attained on Ω_L. So H is nonpositive, and its maximum value on Ω is 0, attained at $(0, 0)$. So H is nonpositive at $(1, 0)$ and at $(0, 1)$, and

$$\psi_1(t)\leqq 0\leqq\psi_2.$$

If $\psi_2 > 0$, the maximum value of H on Ω_L occurs at $(L, 0)$ only, so only $(L, 0)$ is in the support of P_t. Therefore, $\dot{y}(t) = 0$ for almost all t, and $y(t) = y_0$ for all t, which is false. If $\psi_2 = 0$, $\psi_1(t)$ must never be zero, since ψ_0, ψ_1, and ψ_2 never vanish simultaneously. So $\psi_1(t) < 0$ for all t. Then the only point of Ω_L at which H is maximum is $(0, L)$. This implies $x(t) = 0$ for all t, which is false. So the assumption $\psi_0 = 0$ leads to a contradiction, and ψ_0 is -1. Now

$$H(t, x, y, u, v) = -xu^3/[u^2 + v^2] + \psi_1(t)u - \psi_2 v. \tag{16}$$

We next prove $\psi_2 \neq 0$. If $\psi_2 = 0$, let t be any point at which $x(t) > 0$, and let (u^*, v^*) be in the support of P_t. Then H has its maximum value zero at (u^*, v^*), so

$$-x(t)u^{*3}/[u^{*2} + v^{*2}] + \psi_1(t)u^* = 0.$$

If $u^* > 0$, this would imply

$$H(t, x(t), y(t), u^*, v^* + 1) > 0,$$

which is impossible. So $u^* = 0$. Therefore the only point in the support of P_t is $(0, L)$, and $dx/dt = 0$ at almost all points at which $x(t) > 0$. This is incompatible with the end-conditions (14), so the assumption $\psi_2 = 0$ has led to a contradiction. Therefore $\psi_2 > 0$. This, in turn, implies $\psi_1(0) = 0$. For if not, then $\psi_1(0) < 0$, and the function

$$H(0, x(0), y(0), u, v) = \psi_1(0)u - \psi_2 v$$

would assume its maximum value zero only at $(0, 0)$, not at any point in Ω_L. It also implies that $(0, L)$ is not in the support of P_t for any t in $[0, 1]$; for

$$H(t, x(t), y(t), 0, L) = -\psi_2 L < 0.$$

Then

$$\dot{x}(t) = \int_\Omega u P_t(du, dv) > 0$$

for almost all t, so $x(t)$ is strictly increasing.

For $t > 0$, no point (u, v) of Ω with $0 < v < u$ is in the support of P_t. For suppose $0 < v < u$. Define $\theta = v/u$. Then

$$\begin{aligned} 0 = H(t, x(t), y(t), u, v) &= -x(t)u^3/[u^2 + v^2] + \psi_1(t)u - \psi_2 v \\ &= -x(t)u/[1 + \theta^2] + \psi_1(t)u - \psi_2 \theta u, \\ H(t, x(t), y(t), u, 0) &= -x(t)u + \psi_1(t)u, \\ H(t, x(t), y(t), u, u) &= -x(t)u/2 + \psi_1(t)u - \psi_2 u. \end{aligned}$$

From the last two equations,

$$(1 - \theta)H(t, x(t), y(t), u, 0) + \theta H(t, x(t), y(t), u, u) = -x(t)(1 - \theta/2)u + \psi_1(t)u - \psi_2 \theta u.$$

Since

$$(1 - \theta/2) - (1 + \theta^2)^{-1} = -\theta(1 - \theta)^2/2(1 + \theta)^2 < 0,$$

this implies that one of the numbers $H(t, x(t), y(t), u, u)$, $H(t, x(t), y(t), u, 0)$ is positive, in contradiction to the fact that the maximum value of H is zero. So no point (u, v) with $0 < v < u$ is in the support of P_t for any t in $[0, 1]$.

Let A be the set of t in $[0, 1]$ such that $(L, 0)$ is in the support of P_t. If t is in A,

$$0 = H(t, x(t), y(t), L, 0) \geqq H(t, x(t), y(t), L, L),$$

so

$$-x(t)L+\psi_1(t)L-0\geqq -x(t)(L/2)+\psi_1(t)L-\psi_2 L.$$

This implies

$$x(t)\leqq 2\psi_2.$$

If t is in $[0, 1]\backslash A$, the support of P_t contains some point (u, v) with $v\neq 0$, therefore with $0<u\leqq v$. For all t we have by equations (iii) on page 926

$$\psi_1(t)=\int_0^t\left\{\int_\Omega \frac{u^3}{u^2+v^2}P_\tau(du, dv)\right\}d\tau\leqq\int_0^t\left\{\int_\Omega uP_\tau(du, dv)\right\}d\tau=x(t).$$

So if the support of P_t contains (u, v) with $v>0$,

$$\begin{aligned}0=H(t, x(t), y(t), u, v)&=-x(t)u^3/[u^2+v^2]+\psi_1(t)u-\psi_2 v\\&\leqq x(t)\{-u^3/[u^2+v^2]+u\}-\psi_2 v=\{x(t)uv/[u^2+v^2]-\psi_2\}v.\end{aligned}$$

Hence

$$\psi_2\leqq x(t)\{uv/[u^2+v^2]\}\leqq x(t)\{\tfrac{1}{2}\},$$

and therefore

$$x(t)\geqq 2\psi_2.$$

So all points t with $x(t)$ in $[0, 2\psi_2]$ are in A, and all points t with $x(t)$ in $(2\psi_2, 1]$ are in $[0, 1]\backslash A$. For t with $x(t)$ in $[0, 2\psi_2)$, the support of P_t consists of $(L, 0)$ alone, so

$$x(t)=\int_0^t\left\{\int_\Omega uP_\tau(du, dv)\right\}d\tau=\int_0^t L\,d\tau=Lt,$$

$$y(t)=y_0-\int_0^t\left\{\int_\Omega vP_\tau(du, dv)\right\}d\tau=y_0.$$

So if we define $t^*=2\psi_2/L$,

$$x(t^*)=2\psi_2,$$

and $y(t)$ is constantly y_0 on $[0, t^*]$. For $t<t^*$, $(L, 0)$ is in the support of P_t, so

$$H(t, x(t), y(t), L, 0)=0.$$

By continuity,

$$H(t^*, x(t^*), y(t^*), L, 0)=0.$$

For $t>t^*$, the support of P_t is contained in $\{(u, v)$ in Ω: $v\geqq u>0\}$. But when $v\geqq u>0$,

$$\partial^2 H/\partial v^2=2x(t)u^3(u^2+v^2)^{-3}(u^2-3v^2)<0.$$

If the maximum value of H were attained at two different points of Ω_L, there would be two rays $v=m_1u$, $v=m_2u$ $(m_2>m_1\geqq 1)$ on which H and its first partial derivatives vanish. That is,

$$\frac{\partial}{\partial v}H(t, x(t), y(t), 1, m_1)=\frac{\partial}{\partial v}H(t, x(t), y(t), 1, m_2)=0.$$

But this is impossible, since

$$\frac{\partial^2 H}{\partial v^2}(t, x(t), y(t), 1, v)<0\qquad(m_1\leqq v\leqq m_2).$$

Therefore the support of P_t consists of a single point of the arc Ω_L. The optimal pair is an ordinary pair.

We have shown that

$$x(t^*) = \psi_1(t^*) = 2\psi_2,$$

so

$$H(t^*, x(t^*), y(t^*), u, v) = -2\psi_2 u^3/[u^2+v^2] + 2\psi_2 u - \psi_2 v.$$

Therefore

$$\begin{aligned} H(t^*, x(t^*), y(t^*), L, 0) &= 0, \\ H(t^*, x(t^*), y(t^*), L/\sqrt{2}, L/\sqrt{2}) &= 0. \end{aligned} \tag{17}$$

For each point t in $(t^*, 1]$, the support of P_t consists of a single point of Ω_L, which we denote by $(u(t), v(t))$. Let $(\alpha(t), \beta(t))$ be any point that is the limit of $(u(t_n), v(t_n))$ for some sequence of points $t_1, t_2, t_3, \cdots$ of $(t^*, 1]$ tending to t. Then $\beta(t) \geqq \alpha(t)$. By continuity

$$H(t, x(t), y(t), \alpha(t), \beta(t)) = \lim_{n\to\infty} H(t_n, x(t_n), y(t_n), u(t_n), v(t_n)) = 0.$$

But there is only one point $(\alpha(t), \beta(t))$ on Ω_L with $\beta(t) \geqq \alpha(t)$ that satisfies this. If $t > t^*$, $(\alpha(t), \beta(t))$ has to be $(u(t), v(t))$. If $t = t^*$, by (17)

$$\alpha(t) = L/\sqrt{2}, \qquad \beta(t) = L/\sqrt{2}.$$

So the limit of $(u(t), v(t))$ as t tends to t^* from above is $(L/\sqrt{2}, L/\sqrt{2})$.

For t in $(t^*, 1]$, the support of P_t contains a single point, which for almost all t has to be $(\dot{x}(t), \dot{y}(t))$. Since H has its maximum value at this point, its partial derivatives vanish there, so

$$2x(t)\dot{x}(t)^3\dot{y}(t)[\dot{x}(t)^2+\dot{y}(t)^2]^{-2} + \psi_2 = 0. \tag{18}$$

This implies that $\dot{x}$ cannot be zero, so y can be written as a function of x, and (18) amended accordingly. We now have accumulated the following information about the minimizing function, written as $x \to y(x)$:

For $0 \leqq x \leqq 2\psi_2$, $y(x) = y_0$. At $x = 2\psi_2$, there is a corner: the right derivative of y with respect to x is -1. For $x > 2\psi_2$,

$$x = -\psi_2[1+(dy/dx)^2]^2/2[dy/dx]. \tag{19}$$

This is exactly what Newton stated the solution to be.

Parametric equations for the part of the curve to the right of $x = 2\psi_2$ are easily obtained. If we define $\tau = v(t)/u(t)$, by (18) or (19) we find that for almost all t, the function $X(\tau) = x(t(\tau))$ satisfies

$$X(\tau) = (1+\tau^2)^2\psi_2/2\tau. \tag{20}$$

From this,

$$dX/d\tau = (\psi_2/2)[-\tau^{-2}+2+3\tau^2].$$

Let $Y(\tau) = y(t(\tau))$. Since $dY/d\tau = [v/u]\, dX/d\tau = \tau[dX/d\tau]$,

$$dY/d\tau = (\psi_2/2)[-\tau^{-1}+2\tau+3\tau^3],$$

whence

$$Y(\tau) = C + (\psi_2/2)[-\log\tau + \tau^2 + 3\tau^4/4]. \tag{21}$$

We let τ approach 1; then $X(\tau)$ and $Y(\tau)$ approach $x(t^*)=2\psi_2$ and $y(t^*)=y_0$, respectively, and from (21) we obtain

$$C = y_0-(\psi_2/2)(7/4).$$

Substituting this in (21) yields

$$Y(\tau)=y_0+(\psi_2/2)[-\log\tau+\tau^2+3\tau^4/4-7/4].$$

This and (20) are parametric equations for the part of the curve to the right of the straight section $y=y_0$ $(0\leqq x\leqq 2\psi_2)$.

Note added in proof. Professor John Burns has pointed out to me that Newton's least-resistance problem is discussed in *Applied Optimal Control* by A. E. Bryson and Y.-C. Ho [14, pp. 52–55]. Their treatment is of the type we have called "naïve," and also it is not rigorous. But they arrive at Newton's solution, as given above, and they exhibit graphs for three special cases. Also they assert that Newton's formula for the resistance, while inaccurate for subsonic speeds, is very good at speeds above the speed of sound.

REFERENCES

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

[2] G. A. BLISS, *Calculus of Variations*, Mathematical Association of America, Washington, DC, 1925.

[3] E. T. BELL, *Men of Mathematics*, Simon and Schuster, New York, 1937.

[4] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, Chelsea Publishing Co., New York, 1980.

[5] E. J. MCSHANE, *Generalized curves*, Duke Math. J., 6 (1940), pp. 513–536.

[6] ———, *Necessary conditions in generalized-curve problems of the calculus of variations*, Duke Math. J., 7 (1940), pp. 1–27.

[7] ———, *Existence theorems for Bolza problems in the calculus of variations*, Duke Math. J., 7 (1940), pp. 28–61.

[8] J. WARGA, *Necessary conditions for minimum in relaxed variational problems*, J. Math. Anal. Appl., 4 (1963), pp. 129–145.

[9] T. WAZEWSKI, Bull. Acad. Polon. Sci. Ser. Sci. Math. Astron. Phys., 10 (1962), pp. 17–21.

[10] R. V. GAMKRELIDZE, *On sliding optimal states*, Dokl. Akad. Nauk SSR, 143 (1962), pp. 1243–1245. (In Russian.)

[11] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[12] L. D. BERKOVITZ, *Optimal Control Theory*, Appl. Math. Sci., vol. 12, Springer-Verlag, New York, Berlin, 1974.

[13] L. E. EL'GOL'TS, *Differential Equations*, Gordon and Breach, New York, 1961.

[14] A. E. BRYSON AND Y.-C. HO, *Applied Optimal Control: Optimization, Estimation, and Control*, Hemisphere Publishing Co., New York, 1981.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 940-948, September 1989

003

RAPID OSCILLATIONS, CHATTERING SYSTEMS, AND RELAXED CONTROLS*

ZVI ARTSTEIN†

This paper is dedicated to the memory of E. J. McShane.

Abstract. Chattering systems serve as a limit model for control systems with rapidly oscillating control coefficients. This paper investigates the applicability of relaxed controls to chattering systems. In particular, the robustness of optimal relaxed controls and their ordinary approximations are examined. In the case of complete chattering, it is found that the relaxation of the optimal controls can be eliminated.

Key words. relaxed controls, chattering systems, rapid oscillations, robustness

AMS(MOS) subject classifications. 49A10, 49B50

1. Introduction.

An advantage of working with relaxed controls is that under mild conditions, existence of optimal solutions is guaranteed. A justification of coming up with a relaxed solution is that is can usually be approximated well by an ordinary control, namely, the performance is not changed drastically when the optimal relaxed solution is replaced by the approximation. These aspects are explained and demonstrated in the fundamental works of Warga (see [8]) on control systems and Young (see [10]) on the calculus of variations. McShane [7] has developed a general theory of relaxed unbounded controls and has shown when a seemingly optimal relaxed solution is actually an ordinary one.

Robustness is a desired property of optimal solutions and their approximations. Namely, the performance of the chosen control should not be harmed greatly if a small change in the system occurs. A situation where this property may not prevail is when an approximation to a relaxed control is applied to a system with rapidly oscillating coefficients. Indeed, the standard approximations of a relaxed control are rapidly oscillating controls. If the oscillations of the approximating controls are not synchronized with the oscillations of the parameters, a resonance phenomenon may occur, with cost far from optimal. This problem is even more acute when the oscillation rate of the parameters is subject to errors or uncertainties; then the synchronization of the two oscillation rates may become a difficult matter.

The sensitivity to a small change in the highly oscillatory parameters arises not only in connection with relaxed controls. In general, we would like to have a robust model for rapid oscillations, and then analyze the uncertain parameters as small deviations from that model. Such a model has been offered in [1] and [2]. It allows instantaneous oscillations of the parameters; a convergence mode is then defined with which the rapidly oscillating parameters can indeed be treated as deviations from the model. The systems with the infinitely rapid change of parameters are termed *chattering systems.*

In this paper we develop the relaxed controls aspect of the chattering systems. The first interesting phenomenon that we discover is that relaxation can actually be eliminated in case of complete chattering. The reason is that the oscillations of the parameters can be used to generate the oscillations that the good approximations of

* Received by the editors August 10, 1988; accepted for publication (in revised form) December 7, 1988.
† Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel.

the relaxed solutions need; in the limit, ordinary controls suffice. However, there are cases where relaxed controls are needed for optimality. We study them, with their approximations, and examine the robustness property.

We find it best to exhibit the ideas by working out in detail a simple example. This is done in the next section. The abstract theory is treated in § 3, where chattering systems are recalled and relaxed controls for chattering systems are introduced. The elimination of relaxation is examined in § 4. In the closing section we analyze the robustness of relaxed solutions and their ordinary approximations.

2. An example in detail. Consider the following optimal control problem:

$$\begin{aligned} &\text{minimize} && \int_0^1 (x^2(t)+(u^2(t)-1)^2)\,dt, \\ &\text{subject to} && \dot{x}(t)=a_j(t)u(t), \\ &&& x(0)=0. \end{aligned} \tag{2.1}$$

Here the state $x(t)$ and the control $u(t)$ are scalar functions, defined on $[0, 1]$. The index j in the coefficients appears since later on we consider a sequence of problems. For a given control $u(\cdot)$ the integral to be minimized in (2.1) is referred to as the *cost* of using $u(\cdot)$.

We assume that $a_j(t)$ is not identically zero on any subinterval of $[0, 1]$. The infimal cost is then zero, but it cannot be achieved with an ordinary control. To achieve zero cost it is necessary to have $|u(t)|=1$, but then $x(t)$ is not identically zero, and thus zero cost is impossible. To get close to the zero cost, the interval $[0, 1]$ can be divided into k intervals of equal length, and $u_k(t)$ chosen equal to $+1$ or to -1 on alternate intervals; then the cost tends to zero as $k\to\infty$. An optimal relaxed control exists. It is the control, say $v(t)$, which assigns to each t the values $+1$ and -1 with equal probabilities. The sequence $u_k(\cdot)$ constructed earlier converges, indeed, to the optimal relaxed control $v(\cdot)$ as $k\to\infty$ (see Berkovitz [4, § IV.4] or Warga [8, § III.3]).

Consider now a sequence of optimization problems of the type (2.1), each determined by its coefficients $a_j(t)$. We wish to examine systems with rapidly oscillating coefficients; for definiteness we take

$$a_j(t)=\cos 2\pi jt \tag{2.2}$$

and we are interested in the behavior for large j, say $j\to\infty$. An interpretation that we adopt is that j is subject to uncertainty; thus the solution to the optimization problem should be robust with respect to the uncertainty.

As mentioned, the relaxed control $v(t)$ is an optimal solution of (2.1) regardless of the coefficients $a_j(t)$. The performance of the approximations $u_k(t)$ depends strongly, however, on the coefficients. For instance, if $j=k$ the control $u_k(t)$ is a bad replacement of the optimal relaxed control. (If k is fixed, however, and $j\to\infty$, then the approximation is valid. This is not accidental, as we show in the closing section.)

The chattering systems were developed (see [1], [2]) as a robust model for systems with rapidly oscillating coefficients. The abstract framework is recalled in the next section; here we describe the chattering limit of (2.1) as $j\to\infty$. Since the oscillation rate grows indefinitely, in the limit we wish to allow instantaneous oscillations, say at time t_0. This is done by computing the distribution of the values $\cos 2\pi jt$ of the coefficients, in a small neighborhood of t_0, and taking the limit as $j\to\infty$. The limit distribution as the neighborhood shrinks to t_0 depicts the infinitely rapid oscillation at t_0. In the case of (2.1) it is easy to make the computation; the limit distribution

does not depend on t_0 (due to periodicity) and it is given by

$$d\beta = \frac{1}{\pi(1-\sigma^2)^{1/2}}\, d\sigma \tag{2.3}$$

where $\sigma \in [-1, 1]$ and $d\sigma$ is the Lebesgue measure. The interpretation is that in the limit the coefficient $a(t_0)$ "oscillates" at t_0 according to $d\beta$. The control function, in turn, may also oscillate and respond instantaneously to the change of the coefficient. Thus, the control function has the form $u(t, \sigma)$, and the interpretation is that the value $u(t, \sigma)$ is applied when the time is t and the value of the coefficient is σ. The chattering variational problem therefore has the form

$$\begin{aligned} &\text{minimize} && \int_0^1 \left(x^2(t) + \int_{-1}^{1} (u^2(t, \sigma) - 1)^2 \, d\beta \right) dt, \\ &\text{subject to} && \dot{x}(t) = \int_{-1}^{1} \sigma u(t, \sigma)\, d\beta, \\ & && x(0) = 0. \end{aligned} \tag{2.4}$$

By inspection we see that (2.4) has an ordinary optimal solution. Indeed, β is symmetric around 0, and hence the choice $u(t, \sigma) = 1$ for all (t, σ) yields a zero cost. If this optimal solution is applied to the sequence of problems (2.1), the resulting cost tends to zero as $j \to \infty$. (Note that $u(t, \sigma) = 1$ is quite far from the suggested approximations $u_k(t)$, which did not work uniformly anyway.) The phenomenon has a natural explanation as follows. Although the control $u(t, \sigma) = 1$ does not oscillate, the combination of the control with the oscillating parameters $a_j(t)$ provides the dither necessary to mimic the optimal relaxed behavior.

There is also an optimal relaxed control of (2.4); indeed the relaxed control $v(t)$ described earlier is one. It can be interpreted as an optimal relaxed control of the form $v(t, \sigma)$ by letting $v(t, \sigma) = v(t)$. It can be approximated in the topology of relaxed controls by ordinary controls $u_k(t, \sigma)$ in various ways (see Example 3.2, or Warga [8, § III.3]). One possibility is the sequence $u_k(t)$ defined earlier, and indeed, if $u_k(t)$ with large k is fixed, then the cost of applying it to the jth problem is small, as $j \to \infty$.

These observations reflect the properties in the general case analyzed in the rest of the paper.

3. Relaxed controls for chattering systems. First we recall what chattering systems are (following [1]). Then we examine how relaxed controls are applied to the chattering systems.

The ordinary control systems in this paper are of the following form:

$$\begin{aligned} &\text{minimize} && \int_b^a Q(x(t), u(t), t)\, dt, \\ &\text{subject to} && \dot{x} = f(x, t) + g(u, t), \\ & && x(a) = x_0. \end{aligned} \tag{3.1}$$

The chattering system is obtained when at each t the unique control function $g(\cdot, t)$ is replaced by a distribution ζ_t, which is a probability measure on the space G of functions $g(u)$: $R^m \to R^n$ (here $x \in R^n$, $u \in R^m$). The admissible controls, in turn, are functions $u(t, g)$ of both the time t and the coefficient $g(\cdot) \in G$. (We still call g the

coefficient of u, as in (2.1), although in the general case g is a function of u.) The optimization with a chattering system, therefore, has the form

$$\text{minimize} \quad \int_a^b \int_G Q(x(t), u(t, g), t)\zeta_t(dg)\, dt,$$

$$\text{subject to} \quad \dot{x} = f(x, t) + \int_G g(u(t, g))\zeta_t(dg), \tag{3.2}$$

$$x(a) = x_0.$$

Here dg indicates that the integration is done with respect to the variable g; the measure ζ_t depends on the time t. The system (2.4) is an example of type (3.2) with G consisting of $g_\sigma(u) = \sigma u$ and $\sigma \in [-1, 1]$, and $\zeta_t = \beta$ is time invariant. The ordinary system (3.1) is a particular case of (3.2), when ζ_t assigns a unit mass to $g(\cdot, t)$.

We now display the technical assumptions concerning the data.

The function $Q(x, u, t)$, which generates the cost in (3.1) and (3.2), is defined on $R^n \times R^m \times [a, b]$, and it is assumed continuous in the three variables.

Let λ_1 and κ_1 be two positive constants. Let $\mathcal{F}$ denote the collection of mappings $f: R^n \times [a, b] \to R^n$ that are measurable in t and satisfy $|f(x, t)| \leqq \lambda_1(|x|+1)$ and $|f(x, t) - f(y, t)| \leqq \kappa_1|x - y|$. We assume that the function f in (3.2) belongs to $\mathcal{F}$.

Let λ_2 and κ_2 be two positive constants. Let G denote the collection of continuous functions $g(u): R^m \to R^n$ satisfying $|g(u)| \leqq \lambda_2(|u|+1)$ and $|g(u_1) - g(u_2)| \leqq \kappa_2|u_1 - u_2|$. The space G with the topology of uniform convergence on compact sets is compact and metrizable (see, e.g., [1]). Denote by Prob (G) the space of probability measures on G endowed with the weak convergence of measures; the latter is metrizable and Prob (G) is then a compact metric space (see Billingsley [5]). We assume that $t \to \zeta_t$ as a function from $[a, b]$ into Prob (G) is measurable. We denote by $\mathcal{P}$ the space of all such measure-valued functions.

In this paper we work under the following assumption for the sake of simplicity.

Assumption. Let U be a compact set in R^m; all control functions are restricted to have values in U.

With little complication of notation, the set U can be made dependent on t and g. Unbounded controls can also be accounted for by using the techniques of McShane [7] or Warga [8, § VI.4]. We omit the details.

The admissible controls, therefore, are the functions

$$u(t, g): [a, b] \times G \to U, \tag{3.3}$$

which are measurable in both variables.

We now describe how to apply relaxed controls to the chattering systems. Following the ideas of Warga we allow the control function to assign to each (t, g) a probability distribution on U. Thus, if Prob (U) denotes the space of probability distributions on U, then a relaxed control is a mapping

$$v(t, g): [a, b] \times G \to \text{Prob}\,(U) \tag{3.4}$$

assumed measurable in both variables (compare with (3.3)) and where the compact metric structure on Prob (U) is induced by the weak convergence of measures. The ordinary control $u(t, g)$ is a particular case of a relaxed control with the identification of the value $u(t, g)$ with a measure concentrated on $u(t, g)$.

A relaxed control $v(\cdot, \cdot)$ affects (3.2) by averaging for each (t, g) the effects of the points $u \in U$ according to the distribution $v(t, g)$. To employ another integral notation in (3.2) would complicate the formulas. We therefore adopt the notation

introduced in Young [9] (thanks are due to a referee for this reference), and also used in McShane [7], as follows. If $R(u)$ is a function of the variable u, and if $v(t, g)$ is a measure on U, then $\mathfrak{M}(R(v(t, g)))$ denotes the average $\int_U R(u)v(t, g)(du)$. With this notation, the optimization problem (3.2) with the availability of relaxed controls has the following form:

$$\begin{aligned} &\text{minimize} && \int_a^b \int_G \mathfrak{M}(Q(x(t), v(t, g), t))\zeta_t(dg)\, dt, \\ &\text{subject to} && \dot{x} = f(x, t) + \int_G \mathfrak{M}(g(v(t, g)))\zeta_t(dg), \\ & && x(a) = x_0. \end{aligned} \tag{3.5}$$

Existence of optimal solutions to (3.5), and the possibility of approximating them with ordinary controls, follow from the general theory of Warga [8, Chap. 4]. We derive these results from the following similar observations that are used in the sequel.

A relaxed control function $v(\cdot, \cdot)$ induces a measure, say V, on $[a, b] \times G \times U$ as follows. If $[c, d] \subset [a, b]$, $G_0 \subset G$, and $U_0 \subset U$, then

$$V([c, d] \times G_0 \times U_0) = \int_c^d \int_{G_0} v(t, g)(U_0)\zeta_t(dg)\, dt. \tag{3.6}$$

This measure is a $(b-a)$-multiple of a probability measure, since $v(t, g)$ and ζ_t are probability measures. We say that the relaxed controls $v_k(t, g)$ converge to $v_0(t, g)$ if the induced measures V_k converge to V_0 in the weak convergence of measures. With this convergence we have the following.

LEMMA 3.1. *The space of relaxed controls is metric compact; the ordinary controls are dense in it.*

Proof. The space of bounded measures with the weak convergence of measures is metric compact (see Billingsley [5]). Therefore we have only to show that if V_0 is a limit of V_k and the latter are generated by relaxed controls, then so is V_0. But this is a simple disintegration fact, since $\zeta_t \otimes dt$ is fixed in the construction (3.6) and $v_k(t, g)$ are all probability measures. This verifies the compactness. The approximation by ordinary controls can be exhibited by a straightforward construction, since on $[a, b] \times G$ the measure $\zeta_t \otimes dt$ is atomless, as in Berkovitz [4, IV.4] or Warga [8, § III.3].

Example 3.2. An optimal relaxed control for (2.4) is the constant measure-valued control $v(t, \sigma)$ that assigns $+1$ and -1 with equal probabilities. A sequence that converges to it in the topology of relaxed control is the sequence $u_k(t)$ defined in § 2. Another sequence can be constructed by dividing $[-1, 1]$ into k intervals of equal length, and let $u_k(t, \sigma) = +1$ if σ belongs to the even intervals and $u_k(t, \sigma) = -1$ otherwise.

LEMMA 3.3. *If $v_k(\cdot, \cdot)$ converge to $v_0(\cdot, \cdot)$, then the cost of $v_k(\cdot, \cdot)$ converges to the cost of $v_0(\cdot, \cdot)$.*

Proof. To each $v_k(\cdot, \cdot)$ we associate the function

$$\gamma_k(t) = \int_G \mathfrak{M}(g(v_k(t, g)))\zeta_t(dg).$$

The convergence of v_k to v_0 implies (immediately from (3.6)) that $\gamma_k(\cdot)$ converges to $\gamma_0(\cdot)$ in the weak-L_1 convergence on $[a, b]$. A standard continuous dependence result implies that the solutions $x_k(\cdot)$ of the differential equations in (3.5) with v_k converge

uniformly to the solution $x_0(\cdot)$ of the differential equation when v_0 is applied. Define

$$q_k(t) = \int_G \mathfrak{M}(Q(x_k(t), v_k(t, g), t))\zeta_t(dg).$$

The continuity of Q together with the convergence of v_k to v_0 imply that $q_k(\cdot)$ converge weakly in $L_1([a, b])$ to $q_0(\cdot)$. Since the integral of $q_k(\cdot)$ is the cost of v_k, the proof is complete.

THEOREM 3.4. *The optimization problem* (3.5) *has an optimal relaxed control, say* $v_0(\cdot, \cdot)$. *There is a sequence of ordinary controls* $u_k(\cdot, \cdot)$ *which converge to* $v_0(\cdot, \cdot)$ *in the topology of relaxed controls; in particular the cost of* u_k *converges to the infimal cost.*

Proof. The result is a direct consequence of Lemmas 3.1 and 3.3.

4. Elimination of relaxation. If the chattering in the system (3.5) is complete, in a sense to be defined, the ordinary controls yield the same performance as relaxed controls. This is the subject of the present section. The phenomenon is similar to the bang-bang principle (see, e.g., Berkovitz [4]). Here it is the integration with respect to the measures ζ_t that enables the elimination of the relaxation. Indeed, ordinary controls suffice when for each t the measure ζ_t has no atoms (namely, no single point g has positive measure). We call this property complete chattering; the other extreme, no chattering, is the case where for each t the measure ζ_t is concentrated on one point, namely, the ordinary case (3.1).

Before stating the results, we introduce the following convenient tool. Define

$$F(t, g, x) = \{(g(u), Q(x, u, t)): u \in U\}; \tag{4.1}$$

thus $(t, g, x) \to F(t, g, x)$ is a set-valued map with values being subsets of R^{n+1}. The continuity of $g(\cdot)$ and $Q(\cdot, \cdot, \cdot)$, and the topology on G, imply that F has a closed graph. In particular, if $x(t)$ is continuous, then $(t, g) \to F(t, g, x(t))$ has closed values and a closed graph.

We need to integrate the set-valued function F with respect to ζ_t on a subset $G_N(t)$ of G as follows. Let $G_N(t)$ denote the collection of points $g \in G$ such that g is not an atom of ζ_t. The set $G_N(t)$ is then the t-section of the set $G_N = \{(t, g): g$ is not an atom of $\zeta_t\}$. The latter set is measurable with respect to the Lebesgue field on $[a, b]$ and the Borel field on G (see [3, Lemma 4.4]). Denote

$$\Gamma(t, x) = \int_{G_N(t)} F(t, g, x)\zeta_t(dg) \tag{4.2}$$

where the integral is defined to be the set in R^{n+1} of integrals (on $G_N(t)$ with respect to ζ_t) of integrable functions $\gamma(\cdot)$ where $\gamma(g)$ is a selection of F, namely, $\gamma(g) \in F(t, g, x)$ for ζ_t-almost every g. This is the standard integration of set-valued maps (see, e.g., Klein and Thompson [6]). Here t and x are parameters of the integration; parametrized integration of set-valued maps was examined in [3], and the main result of the latter is the tool we use in the sequel.

The result we state and prove is somewhat more general than the one promised in the beginning of the section; the latter follows as a simple consequence.

THEOREM 4.1. *Let* $v_0(t, g)$ *be a relaxed control for* (3.5). *Let* $x_0(t)$ *be the trajectory generated by* v_0 *and let* c_0 *be its cost. There exists a relaxed control* $v_1(t, g)$ *with the same cost, that generates the same trajectory, and that has the property that whenever* g *is not an atom of* ζ_t, $v_1(t, g)$ *is concentrated on one point.*

Proof. Denote

$$\theta_0(t, g) = \mathfrak{M}(g(v_0(t, g)), Q(x_0(t), v_0(t, g), t)), \tag{4.3}$$

namely, $\theta_0(t, g)$ is the integrand in the problem (3.5) when $v_0(t, g)$ is applied to it. Clearly, $\theta_0(t, g)$ is in co $F(t, g, x_0(t))$, the convex hull of $F(t, g, x_0(t))$. Denote

$$\gamma_0(t) = \int_{G_N(t)} \theta_0(t, g)\zeta_t(dg). \tag{4.4}$$

Then $\gamma_0(t) \in \text{co}\,\Gamma(t, x_0(t))$ (see, e.g., Klein and Thompson [6, Thm. 18.1.9]). Since ζ_t is atomless on $G_N(t)$, the set $\gamma(t, x(t))$ is actually convex [6, Cor. 18.1.10], and it follows that $\gamma_0(t) \in \Gamma(t, x_0(t))$. By the theorem in Artstein [3], it follows that there exists a measurable selection $\theta_1(t, g)$ of $F(t, g, x_0(t))$ such that

$$\int_{G_N(t)} \theta_1(t, g)\zeta_t(dg) = \gamma_0(t). \tag{4.5}$$

A standard implicit functions lemma (see, e.g., Berkovitz [4, Thm. 7.1]) implies the existence of an ordinary control $u_1(t, g)$ defined on G_N such that

$$\theta_1(t, g) = (g(u_1(t, g)), Q(x_0(t), u_1(t, g), t)).$$

Consider the relaxed control $v_1(t, g)$ that is equal to $v_0(t, g)$ if g is an atom of ζ_t, and equal to the ordinary control $u_1(t, g)$ if g is not an atom of ζ_t. By (4.3)-(4.5) it follows that $x_0(t)$ is also the trajectory generated by $u_1(t, g)$ and, given that, it follows that v_0 and v_1 have the same cost. This completes the proof.

COROLLARY 4.2. *Suppose that for every t the measure ζ_t is atomless. Let $v_0(t, g)$ be a relaxed control, let $x_0(t)$ be the trajectory generated by it, and let c_0 be its cost. Then there exists an ordinary control $u_0(t, g)$ that generates the same trajectory and has the same cost. In particular, the problem* (3.2) *then has an optimal ordinary solution.*

Proof. The first part is a particular case of Theorem 4.1; the conclusion then follows from Theorem 3.4.

5. Robustness. The control policy that we recommend when rapid oscillations with uncertainty are present is as follows. First solve the limit chattering problem. Then apply the solution to the case of rapid oscillations. Success of such a plan is the robustness in the title of the section. It holds, however, only under some restrictions on the control functions that are used. In this section we determine these restrictions and establish the existence of ordinary controls satisfying these restrictions. Results in this direction appear in [1, § 7] for ordinary controls; here we complement these results and examine in particular the relaxed controls case.

We recall the limit notion that we use to determine if a problem of type (3.2), or (3.1), is a small perturbation of another problem. Note that each problem is characterized by a pair $(f(x, t), \zeta_t) \in \mathcal{F} \times \mathcal{P}$ (see § 3).

Let $f_j, j = 0, 1, 2, \cdots$, belong to $\mathcal{F}$. We say that f_j converge to f_0 if for every $x \in R^n$ and $t \in [a, b]$ the sequence $\int_a^t f_j(x, s)\,ds$ converges to $\int_a^b f_0(x, s)\,ds$. This is the standard convergence that ensures continuous dependence of solutions. The topology induced by it on $\mathcal{F}$ is metric and compact (see [1]).

We now define a convergence on $\mathcal{P}$. First we identify the measure-valued map $t \to \zeta_t$ with the measure $\boldsymbol{\zeta} = \int_a^b \zeta_t \otimes dt$ on $[a, b] \times G$, namely,

$$\boldsymbol{\zeta}(D) = \int_a^b \zeta_t(D_t)\,dt$$

when $D \subset [a, b] = G$ and D_t is the t-section of D. Then $\boldsymbol{\zeta}$ is a $(b-a)$-multiple of a probability measure. With this identification, the convergence on $\mathcal{P}$ is then the weak convergence of measures (see, e.g., Billingsley [5]). With this convergence $\mathcal{P}$ becomes metric and compact (see [1]).

We need the following notation. By cost (v, f, ζ) we denote the cost of the control v when applied to the problem generated by $(f, \zeta) \in \mathcal{F} \times \mathcal{P}$. The first robustness result is under strong conditions on the controls, as follows.

PROPOSITION 5.1. *Let $v_0(t, g)$ be an admissible relaxed control that is continuous in the g variable. Let (f_j, ζ_j) converge to (f_0, ζ_0) in $\mathcal{F} \times \mathcal{P}$. Then* cost (v_0, f_j, ζ_j) *converge to* cost (v_0, f_0, ζ_0).

Proof. Proposition 7.1 of [1] is the analogous result for ordinary controls; the proof is similar. Here is the key step. Let $\gamma_j(t)$ be defined by (4.3)–(4.4) above, when v_0 is the control and (f_j, ζ_j) the data of the problem. The continuity of $v_0(t, \cdot)$ and the convergence on $\mathcal{F} \times \mathcal{P}$ imply that $\gamma_j(\cdot)$ converge to $\gamma_0(\cdot)$ in the weak-L_1 sense on $[a, b]$. This implies the convergence of the cost.

The continuity condition may be severe in some situations; for instance, if U is disconnected (say U contains a finite number of points) and when we look for ordinary controls. (Note, however, that both the optimal relaxed solution and the optimal ordinary solution in (2.4) are continuous, with $U = \{-1, 1\}$.) The following result establishes only a near robustness result, but under eased conditions, which, as we see later, can be fulfilled in general.

PROPOSITION 5.2. *Let $K \subset [a, b] \times G$ be a compact set such that $\zeta_0(([a, b] \times G) \setminus K) < \delta$. Let $v_0(t, g)$ be an admissible relaxed control such that $(t_0, g_0) \in K$ implies that $v_0(t_0, \cdot)$ is continuous at the point g_0. Suppose (f_j, ζ_j) converge to (f_0, ζ_0) in $\mathcal{F} \times \mathcal{P}$. Then* limsup $|\text{cost}\,(v_0, f_j, \zeta_j) - \text{cost}\,(v_0, f_0, \zeta_0)| \leqq \varepsilon(\delta)$, *with $\varepsilon(\delta) \to 0$ as $\delta \to 0$.*

Proof. Define $v_1(t, g)$ such that $v_1(t, g) = v_0(t, g)$ if $(t, g) \in K$ and $v_2(t, g)$ is continuous in the variable g. This can be done in a standard way, since the space of probability measures on U is convex. By the previous proposition cost (v_1, f_j, ζ_j) converges to cost (v_1, f_0, ζ_0). The result would then follow if we prove that $|\text{cost}\,(v_1, f_0, \zeta_0) - \text{cost}\,(v_0, f_0, \zeta_0)| \leqq \varepsilon(\delta)$ with $\varepsilon(\delta) \to 0$ as $\delta \to 0$. But this follows directly from (3.5). Indeed, the continuous dependence implies that the trajectory $x_1(t)$ generated by v_1 is close to the trajectory $x_0(t)$ generated by v_0; the continuity of $Q(\cdot, u, t)$ and the fact that the cost is an integral on $[a, b] \times G$ with respect to ζ_0, and $([a, b] \times G) \setminus K$ has ζ_0 measure δ, imply the result.

Optimal solutions of (3.2) may not satisfy the robustness or the near robustness conditions of the previous results. We wish, therefore, to establish the existence of approximate solutions that satisfy the conditions. The following result verifies the existence of a nearly robust approximate solution in the general case, and a robust approximate solution under an additional condition on U. In particular, since the optimal solutions may be relaxed, we wish to verify the existence of robust, or nearly robust, ordinary controls.

Denote by $\rho(v_1, v_2)$ the distance between the control functions $v_1(t, g)$ and $v_2(t, g)$; namely, the metric that generates the convergence of control functions (see Lemma 3.1). Recall that the cost is continuous with respect to this metric.

THEOREM 5.3. *Let $v(t, g)$ be an admissible control for the problem* (3.2) *generated by the data (f, ζ). Let $\varepsilon > 0$ and $\delta > 0$ be given. There exists an ordinary control function $u(t, g)$ such that $\rho(v, u) < \varepsilon$ and such that a compact set $K \subset [a, b] \times G$ exists, with $\zeta(([a, b] \times G) \setminus K) < \delta$ and $u(t_0, \cdot)$ is continuous at points g_0 with $(t_0, g_0) \in K$. If, in addition, U is a convex set, then K can be chosen equal to $[a, b] \times G$.*

Proof. The ordinary controls are dense in the space of relaxed controls (Lemma 3.1), therefore an admissible, namely measurable, ordinary control $u_1(t, g)$ exists such that $\rho(v, u_1) < \varepsilon/2$. Given $\delta_1 > 0$, by the Lusin Theorem (Warga [8, p. 70]) there exists a $K \subset [a, b] \times G$ such that $\zeta(([a, b] \times G) \setminus K) < \delta_1$ and u_1 restricted to K is continuous. The number δ_1 can be chosen such that $\delta_1 \leqq \delta$ and such that any extension u of the

restriction of u_1 to K, to the space $[a, b] \times G$, will satisfy $\rho(u_1, u) < \varepsilon/2$. What we should show now is the existence of such an extension, with the desired continuity properties. If U is convex, then, since K is compact, a standard extension technique provides a continuous extension u of the restriction of u_1 to K, which is the second claim of the theorem. If U is not convex, we proceed as follows. To each $(t, g) \notin K$ we associate the subset $K(t, g)$ of K, consisting of all points in K closest in the space $[a, b] \times G$ to (t, g). Then $K(t, g)$ is a multifunction with a closed graph and nonempty values. It has a measurable selection (e.g., Klein and Thompson [6, Chap. 14]), say $r(t, g)$. Define $u(t, g)$ by $u(t, g) = u_1(t, g)$ if $(t, g) \in K$ and $u(t, g) = u_1(r(t, g))$ otherwise. Then u is measurable, and continuous at points (t, g) in K. This completes the proof.

Clearly, without the convexity of U there may not exist an ordinary control $u(t, g)$ near the optimal control, and that is continuous in g. Take for instance the differential equation in (2.2) and (2.4), with $Q(x, u, t) = x$, and $U = \{-1, 1\}$. The optimal control is $u(t, \sigma) = -\operatorname{sgn} \sigma$, and it cannot be approximated by a control continuous in σ.

A conclusion of the previous discussion is that near (in the topology of relaxed controls) every optimal relaxed solution there exists an ordinary solution that is either robust or nearly robust. However, note that this robustness is not uniform, as we have seen in the example of § 2. When u_k is fixed near the optimal relaxed control v, then for $j \to \infty$ only a small error is guaranteed. But with a varying index k, there is no small bound on the errors as $j \to \infty$.

REFERENCES

[1] Z. ARTSTEIN, *A variational convergence that yields chattering systems*, Ann. Inst. H. Poincaré, Anal. Non Linéare, to appear.

[2] ———, *Chattering linear systems: A model for rapidly oscillating coefficients*, Math. Control Signals Systems, to appear.

[3] ———, *Parametrized integration of multifunctions with applications to control and optimization*, SIAM J. Control Optim., 27 (1989), to appear.

[4] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[5] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[6] E. KLEIN AND A. C. THOMPSON, *Theory of Correspondence*, Wiley-Interscience, New York, 1984.

[7] E. J. MCSHANE, *Relaxed controls and variational problems*, SIAM J. Control Optim., 5 (1967), pp. 438-485.

[8] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[9] L. C. YOUNG, *Generalized curves and the existence of an attained absolute minimum in the calculus of variations*, C. R. Sci. Lett. Varsovie CIII, 30 (1937), pp. 212-234.

[10] ———, *Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 949-975, September 1989

004

OBSERVABILITY OF SYSTEMS UNDER UNCERTAINTY*

JEAN-PIERRE AUBIN† AND HALINA FRANKOWSKA†

This paper is dedicated to the memory of E. J. McShane.

Abstract. The evolution of the state $x(\cdot)$ of a system under uncertainty governed by a differential inclusion

$$\text{for almost all } t \in [0, T], \quad x'(t) \in F(t, x(t))$$

is observed through an observation map H:

$$\forall t \in [0, T], \quad y(t) \in H(x(t)).$$

The set-valued character due to the uncertainty leads to the introduction of the following:

Sharp input-output map, which is the (usual) product

$$\forall x_0 \in X, \quad I_-(x_0) := (H \circ \mathcal{S})(x_0) := \bigcup_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot)).$$

Hazy input–output map, which is the **square product**

$$\forall x_0 \in X, \quad I_+(x_0) := (H \square \mathcal{S})(x_0) := \bigcap_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot)).$$

Where $\mathcal{S}$ denotes the solution map, recovering the input x_0 from the outputs $I_-(x_0)$ or $I_+(x_0)$ means that these input-output maps are "injective" in the sense that, locally,

$$x_1 \neq x_2 \Rightarrow I(x_1) \cap I(x_2) = \emptyset.$$

Criteria for both sharp and hazy local observability are provided in terms of (global) sharp and hazy observability of the variational inclusion

$$w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t)),$$

which is a "linearization" of the differential inclusion along a solution $\bar{x}(\cdot)$, where for almost all t, $DF(t, x, y)(u)$ denotes the contingent derivative of the set-valued map $F(t, \cdot, \cdot)$ at a point (x, y) of its graph. These conclusions are reached by implementing the following strategy:

1. Provide a general principle of local injectivity and observability of a set-valued map I, which derives these properties from the fact that the kernel of an adequate derivative of I is equal to zero.
2. Supply chain rule formulas that allow computation of the derivatives of the usual product I_- and the square product I_+ from the derivatives of the observation map H and the solution map $\mathcal{S}$.
3. Characterize the various derivatives of the solution map $\mathcal{S}$ in terms of the solution maps of the associated variational inclusions.
4. Piece together these results for deriving local sharp and hazy observability of the original system from sharp and hazy observability of the variational inclusions.
5. Study global sharp and hazy observability of the variational inclusions.

Key words. convex process, set-valued derivative, differential inclusion, inverse mapping theorem, observability, uncertain system, variational inclusion

AMS(MOS) subject classifications. 93B07, 93C10

1. Introduction.

We describe the evolution $t \in [0, T] \mapsto x(t) \in X$ of the state $x(\cdot)$ of a system under uncertainty by a differential inclusion

$$\text{for almost all } t \in [0, T], \quad x'(t) \in F(t, x(t)) \tag{1}$$

where the set-valued map takes into account disturbances and/or perturbations of the

* Received by the editors December 7, 1987; accepted for publication (in revised form) November 21, 1988.

† CEREMADE, Université de Paris-Dauphine, Paris, France and International Institute for Applied Systems Analysis, Laxenburg, Austria.

system. Let us mention a familiar representation of uncertainty:

$$\text{for almost all } t \in [0, T], \quad x'(t) = f(t, x(t)) + g(t, d(t)), \quad d(t) \in D(t).$$

This system is observed through an observation map H that generally is a set-valued map from the state space X to some observation space Y, that associates with each solution to the differential inclusion (1) an observation $y(\cdot)$ satisfying

$$\forall t \in [0, T], \quad y(t) \in H(x(t)). \tag{2}$$

For instance, y may be given in a parametrized form:

$$\forall t \in [0, T], \quad y(t) = h(x(t)) + \varepsilon(t), \quad \varepsilon(t) \in Q(t).$$

We assume for simplicity that H does not depend on the time t, but we shall provide in the appropriate remarks the extensions to the time-dependent case.

Observability concepts deal with the possibility of recovering the initial state $x_0 = x(0)$ of the system knowing only the evolution of an observation $t \in [0, T] \mapsto y(t)$ during the interval $[0, T]$, and naturally, knowing the laws (1) and (2). Once we get the initial state x_0, we may, by studying the differential inclusion, gather information about the solutions starting from x_0, using many results provided by the theory of differential inclusions. For instance, under an adequate Lipschitz property, we know that for every $\bar{x}(\cdot) \in \mathcal{S}(x_0)$,

$$\mathcal{S}(x_0) \in \bar{x}(\cdot) + M \int_0^T \operatorname{diam}(F(t, \bar{x}(t)))\, dt\, B$$

where $\mathcal{S}(x_0)$ denotes the set of all trajectories of (1) starting at x_0, M is a constant independent of $\bar{x}(\cdot)$ and B denotes the closed unit ball in the Sobolev space $W^{1,1}(0, T)$.

Let $\mathcal{S} := \mathcal{S}_F$ from X to $\mathscr{C}(0, T; X)$ denote the solution map associating with every initial state $x_0 \in X$ the (possibly empty) set $\mathcal{S}(x_0)$ of solutions to the differential inclusion (1) starting at x_0 at the initial time $t = 0$.

In other words, we have introduced an input-output system where the

1. inputs are the **initial states** x_0, and the

2. outputs are the **observations** $y(\cdot) \in H(x(\cdot))$ of the evolution of the state of the system through H:

$$\begin{array}{ccccc}
\textbf{Inputs} & \overset{\mathcal{S}}{\rightsquigarrow} & \textbf{States} & \overset{H}{\mapsto} & \textbf{Outputs} \\
\downarrow & & \downarrow & & \downarrow \\
X \ni x_0 & \rightsquigarrow & x(\cdot) \in \mathcal{S}(x_0) & \mapsto & y(\cdot) \in H(x(\cdot)) \\
\uparrow & & \uparrow & & \uparrow \\
\text{Initial states} & & \begin{cases} x'(t) \in F(t, x(t)) \\ x(0) = x_0 \end{cases} & & \text{Observations.}
\end{array}$$

It remains to define an input-output map. But, because of the set-valued character (the presence of uncertainty), we can conceive two dual ways for defining composition products of the set-valued maps $\mathcal{S}$ from X to the space $\mathscr{C}(0, T; X)$ and H from $\mathscr{C}(0, T; X)$ to $\mathscr{C}(0, T; Y)$. So, for systems under uncertainty, we have to deal with **two input–output maps** from X to $\mathscr{C}(0, T; Y)$:

Sharp input–output map, which is the (usual) product

$$\forall x_0 \in X, \quad I_-(x_0) := (H \circ \mathcal{S})(x_0) := \bigcup_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot)).$$

Hazy input–output map, which is the **square product**

$$\forall x_0 \in X, \quad I_+(x_0) := (H \square \mathcal{S})(x_0) := \bigcap_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot)).$$

The **sharp** input-output map tracks the evolution of **at least one** state starting from some initial condition x_0 whereas the **hazy** input-output map tracks **all** such evolutions.

Opinions may differ about which would be the "right" input-output map, just because they depend on the context in which a given problem is stated. So, we shall study observability properties of **both** the sharp and hazy input-output maps.

Recovering the input x_0 from the outputs $I_-(x_0)$ or $I_+(x_0)$ means that the set-valued maps are "injective" in some sense.

When H and $\mathcal{S}$ are single-valued maps, the input-output map is called observable whenever the product $I := H \circ \mathcal{S}$ is injective, i.e.,

$$H\mathcal{S}(x_1) = H\mathcal{S}(x_2) \Rightarrow x_1 = x_2. \tag{3}$$

When we adapt this definition to the set-valued case, we come up with two possibilities: If I stands now for either I_- or I_+, we can require either the property

$$I(x_1) = I(x_2) \Rightarrow x_1 = x_2$$

or the stronger condition

$$I(x_1) \cap I(x_2) \neq \emptyset \Rightarrow x_1 = x_2.$$

The first way would not be, in general, useful in the framework of uncertain systems since we often observe just one output $y(\cdot) \in H\mathcal{S}(x_0)$ and not the whole set of possible outputs $H\mathcal{S}(x_0)$. That is why we will adopt the second point of view, by saying that the sharp or hazy input-output map I is "observable" if

$$x_1 \neq x_2 \Rightarrow I(x_1) \cap I(x_2) = \emptyset. \tag{4}$$

If this property holds only on a neighborhood of some x_0, we shall say that I is "locally observable" around x_0.

This is a very pleasant concept that we will study for hazy input-output maps.

However, it is a bit too strong for sharp observability, and we will be content with the weaker condition that the inverse image $I^{-1}(y_0)$ of some observation y_0 contains **at most** one input x_0:

$$x_1 \neq x_0 \Rightarrow y_0 \notin I(x_1). \tag{5}$$

If this is the case, we will say that the input-output map I is "observable" **at** (x_0, y_0) and "locally observable" **at** (x_0, y_0) if it holds only on a neighborhood of x_0.

In other words, sharp observability at $(x_0 y_0)$ means that whenever y_0 is an observation of some solution $x^*(\cdot)$, i.e., $y_0 \in H(x^*(\cdot))$, then $x^*(0) = x_0$. Local sharp observability means that the above holds true only for those $x^*(0)$ not too far from x_0.

Hazy observability at (x_0, y_0) means that y_0 can be a "common" observation only for one input x_0. In other words, if we (hopefully) observe an output y_0, which is a common observation of all solutions $x(\cdot) \in \mathcal{S}(\bar{x}_0)$, then $\bar{x}_0 = x_0$.

Actually, the purpose of this paper is to derive local observability of both the sharp and hazy input-output maps from the global sharp and hazy observability at zero of "variational inclusions" through a linearization of the input-output map. The linearization techniques based on the differential calculus and inverse function theorems for set-valued maps has been successfully used in the study of local controllability of differential inclusions and control systems with feedbacks. (See [12], [13], [10], [11], [20].)

Here, variational inclusions are "linearizations" of the differential inclusion (1) along a solution $\bar{x}(\cdot) \in \mathcal{S}(x_0)$ of the form

$$w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \tag{6}$$

where for almost all t, $DF(t, x, y)(u)$ denotes an adequate concept of derivative (the contingent derivative, defined below) of the set-valued map $F(t, \cdot, \cdot)$ at a point (x, y) of its graph. Let us just say for the time that they are set-valued analogues of continuous linear operators.

(These linearized differential inclusions are called **variational inclusions** because they extend (in various ways) the classical variational equations of ordinary differential equations: their solutions starting at some u provide the directional derivative of the solution to the initial system in the direction u.)

To say that the variational inclusion is **hazily** (respectively, **sharply**) observable at zero amounts to saying that whenever **all** (respectively, **at least one**) solutions $w(\cdot)$ to the variational inclusion (6) starting at u satisfy

$$\forall t \in [0, T], \quad H'(\bar{x}(t))w(t) = 0 \tag{7}$$

then $u = 0$.

To reach such conclusions, we shall choose the following strategy:

1. Provide a general principle of local injectivity and observability of a set-valued map I that derives these properties from the fact that the kernel of an adequate derivative of I is equal to zero.
2. Supply chain rule formulas that allow computing the derivatives of the usual product I_- and the square product I_+ from the derivatives of the observation map H and the solution map $\mathcal{S}$.
3. Characterize the various derivatives of the solution map $\mathcal{S}$ in terms of the solution maps of the associated variational inclusions.
4. Piece together these results for deriving local sharp and hazy observability of the original system from sharp and hazy observability of the variational inclusions.
5. Study global sharp and hazy observability of the variational inclusions. (This has already been done in [5], for time-independent closed convex processes, where it was shown that sharp observability is a dual concept of controllability and where various characterizations were provided. See the last section for the comments on the observability of a system around an equilibrium).

But, before implementing this program, we have to avoid the trivial case when the hazy input-output map I_+ takes (locally) empty values.

For doing that, we "project" the differential inclusion (1) onto a differential inclusion

$$\text{for almost all } t \in [0, T], \quad y'(t) \in G(t, y(t)) \tag{8}$$

in such a way that the following property holds true:

$$\begin{aligned} &\forall (x_0, y_0) \in \text{Graph}\,(H) \text{ all solutions } x(\cdot) \text{ to (1) and } y(\cdot) \text{ to (8) satisfy} \\ &\forall t \in [0, T], y(t) \in H(x(t)). \end{aligned} \tag{9}$$

If such is the case, then the hazy input-output map I_+ is well defined.

To proceed further, we need to introduce the concept of "contingent derivative" of a set-valued map H from a Banach space X to a Banach space Y at a point (x, y)

of its graph: It is the set-valued map $DH(x, y): X \rightsquigarrow Y$ that associates with any direction u the set $DH(x, y)(u)$ of directions v satisfying

$$\liminf_{h \to 0+,\, u' \to u} d\left(v, \frac{H(x+hu')-y}{h}\right) = 0. \tag{10}$$

The choice of this particular derivative is motivated by the fact that its graph is the **contingent cone** to the graph of H at (x, y), where the contingent cone $T_K(x)$ to $K \subset X$ at $x \in K$ is the set of directions $v \in X$ such that

$$\liminf_{h \to 0} d(x+hv, K)/h = 0.$$

For our purpose, the contingent cone plays a major role compared to other tangent cones. However, we shall need other tangent cones and associated derivatives.

The map H is said to be "derivable" if for every (x, y) in the graph of H, v belongs to $DH(x, y)(u)$ if and only if

$$\lim_{h \to 0+} d\left(v, \frac{H(x+hu)-y}{h}\right) = 0.$$

We extend the concept of $\mathscr{C}^1$-function by saying that H is "sleek" if and only if

$$\text{Graph}\,(H) \ni (x, y) \rightsquigarrow \text{Graph}\,(DH(x, y)) \text{ is lower semicontinuous}$$

where $\rightsquigarrow$ means "maps to." (It underlines the set-valued character of the map under consideration.) In this case, the graph of $DH(x, y)$ is a closed convex cone. Maps whose graphs are closed convex cones, called closed convex processes, are the set-valued analogues of continuous linear operators, and enjoy most of their properties.

Returning to the projection problem, we shall say that a set-valued map $G: [0, T] \times Y \rightsquigarrow Y$ is a "Lipschitzian square projection" of the set-valued map $F: [0, T] \times X \rightsquigarrow X$ by H if and only if

$$\begin{aligned} &\text{(i)} \quad F \times G \text{ is Lipschitzian around } [0, T] \times \text{Graph}\,(H), \\ &\text{(ii)} \quad \forall (x, y) \in \text{Graph}\,(H),\ G(t, y) \subset \bigcap_{v \in F(t,x)} DH(x, y)(v). \end{aligned} \tag{11}$$

In this paper "F is Lipschitzian on $K \subset [0, T] \times X$" means that for all (t, x), $(t, y) \in K$

$$F(t, x) \subset F(t, y) + k(t)\|x-y\|B,$$

where $k \in L^1(0, T)$. We shall prove that *if there exists a Lipschitzian square projection of F by H, then the hazy input-output map $I_+ := H \square \mathcal{S}$ has nonempty values for any initial value $y_0 = H(x_0)$.*

We state now the observability properties of the hazy input-output map around a solution $\bar{x}(\cdot)$ to the differential inclusion (1). We assume that F satisfies the following assumptions:

$$\begin{aligned} &\text{(i)} \quad \forall x \in X \text{ the set-valued map } F(\cdot, x) \text{ is measurable,} \\ &\text{(ii)} \quad \forall t \in [0, T],\ \forall x \in X,\ F(t, x) \text{ is a closed nonempty set,} \\ &\text{(iii)} \quad \exists k(\cdot) \in L^1(0, T) \text{ such that for almost all } t \in [0, T] \\ &\qquad \text{the map } F(t, \cdot) \text{ is } k(t)\text{-Lipschitzian.} \end{aligned} \tag{12}$$

THEOREM 1.1. *Let us assume that H is continuously differentiable, that F satisfies assumptions* (12), *that it has linear growth in the sense that*

$$\exists c > 0 \quad \textit{such that } \|F(t, x)\| := \sup_{y \in F(t,x)} \|y\| \leqq c(\|x\|+1)$$

and that is has a Lipschitzian square projection G by H.

1) *If F is derivable and if for some $\bar{x}(\cdot) \in \mathcal{S}(x_0)$ the contingent variational inclusion*

(13) *for almost all $t \in [0, T]$, $w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t))$*

is globally hazily observable through $H'(\bar{x}(\cdot))$ at zero, then the system (1) *is locally hazily observable through H* **at** *$(x_0, H(\bar{x}))$.*

2) *If F is sleek and if for every solution $x(\cdot)$ to the differential inclusion* (1) *starting at x_0, the contingent variational inclusion*

(14) *for almost all $t \in [0, T]$, $w'(t) \in DF(t, x(t), x'(t))(w(t))$*

is globally hazily observable through $H'(x(\cdot))$ at 0, then the system (1) *is locally hazily observable through H* **around** *x_0.*

Observability properties of sharp input-output maps require stronger assumptions. We state first the result for the simpler, convex case.

THEOREM 1.2. *Let us assume that H is linear and that the graphs of the set-valued maps $F(t, \cdot): X \rightsquigarrow X$ are closed and convex. If for some $\bar{x}(\cdot) \in \mathcal{S}(x_0)$ the contingent variational inclusion* (13) *is globally sharply observable through H at zero, then the system* (1) *is globally sharply observable through H at $(x_0, H(\bar{x}))$.*

A more general case requires some additional assumptions.

THEOREM 1.3. *Assume that F has closed convex images, is continuous, derivable, Lipschitz in the second variable with a constant independent of t and that the growth of F is linear with respect to the state. Let H be a twice continuously differentiable function from X to another finite-dimensional vector-space Y. Consider an observation $y^* \in I_-(x_0)$ and assume that for every solution $\bar{x}(\cdot)$ to the differential inclusion* (1) *satisfying $y^*(\cdot) = H(\bar{x}(\cdot))$ and for all $t \in [0, T]$ we have*

$$\operatorname{Ker} H'(\bar{x}(t)) \cap [F(t, \bar{x}(t)) - F(t, \bar{x}(t))]^{\perp} = \{0\}.$$

If for all $\bar{x}(\cdot)$ as above the contingent variational inclusion (13) *is globally sharply observable through $H'(\bar{x}(\cdot))$ at 0, then the system* (1) *is locally sharply observable through H at (x_0, y^*).*

2. Hazy and sharp input–output systems. Let us consider a set-valued input-output system of the following form built through a differential inclusion

(15) for almost all $t \in [0, T]$, $x'(t) \in F(t, x(t))$

whose dynamics are described by a set-valued map F from $[0, T] \times X$ to X, where X is a finite-dimensional vector-space (the **state space**) and $0 < T \leqq \infty$. It governs the (uncertain) evolution of the state $x(\cdot)$ of the system. The **inputs** are the **initial states** x_0 and the **outputs** are the **observations** $y(\cdot) \in H(x(\cdot))$ of the evolution of the state of the system through a single-valued (or set-valued) map H from X to an **observation space** Y.

Let $\mathcal{S} := \mathcal{S}_F$ from X to $\mathcal{C}(0, T; X)$ denote the solution map associating with every initial state $x_0 \in X$ the (possibly empty) set $\mathcal{S}(x_0)$ of solutions to differential inclusion (15) starting at x_0 at the initial time $t = 0$.

We can conceive two dual ways for defining composition products of set-valued maps G from a Banach space X to a Banach space Y and a set-valued map H from Y to a Banach space Z (which naturally coincide when H and G are single-valued).

DEFINITION 2.1. Let X, Y, Z be Banach spaces and $G: X \rightsquigarrow Y$, $H: Y \rightsquigarrow Z$ be set-valued maps.

1. The usual composition product (called simply the **product**) $H \circ G: X \rightsquigarrow Z$ of H and G at x is defined by

$$(H \circ G)(x) := \bigcup_{y \in G(x)} H(y).$$

2. The **square product** $H \square G: X \rightsquigarrow Z$ of H and G at x is defined by

$$(H \square G)(x) := \bigcap_{y \in G(x)} H(y).$$

Remarks. 1. The observability problems that we address involve the inversion of these input-output maps.

There are two ways to adapt to the set-valued case the formula that states that the inverse of a product is the product of the inverses (in reverse order), since we know that there are two ways of defining the inverse image by a set-valued map $\mathcal{S}$ of a subset M:

$$\text{(a)} \quad \mathcal{S}^-(M) := \{x \mid \mathcal{S}(x) \cap M \neq \varnothing\},$$

$$\text{(b)} \quad \mathcal{S}^+(M) := \{x \mid \mathcal{S}(x) \subset M\}.$$

We then observe the following formulas of the inverse of composition products:

$$\text{(i)} \quad (H \circ \mathcal{S})^{-1}(y) = \mathcal{S}^-(H^{-1}(y)),$$

$$\text{(ii)} \quad (H \square \mathcal{S})^{-1}(y) = \mathcal{S}^+(H^{-1}(y)).$$

This may provide a further justification of the introduction of these two "dual" composition products.

2. Recall also that a set-valued map $\mathcal{S}$ is upper semicontinuous if and only if the inverse images $\mathcal{S}^+$ of open subsets are open and that it is lower semicontinuous if and only if the inverse images $\mathcal{S}^-$ of open subsets are open.

3. Observe finally that square products are implicitly involved in the factorization of maps. Let X be a subset, $\mathcal{R}$ be an equivalence relation on X and ϕ denote the canonical surjection from X onto the factor space $X/\mathcal{R}$. If f is a single-valued map from X to Y, its factorization $\tilde{f}: X/\mathcal{R} \mapsto Y$ is defined by

$$\tilde{f}(\xi) := (f \square \phi^{-1})(\xi).$$

It is nontrivial if and only if f is consistent with the equivalence relation $\mathcal{R}$, i.e., if and only if $f(x) = f(y)$ whenever $\phi(x) = \phi(y)$.

When $F: X \rightsquigarrow Y$ is a set-valued map, we can define its factorization $\tilde{F}: X/\boldsymbol{R} \rightsquigarrow Y$ by

$$\tilde{F}(\xi) := (F \square \phi^{-1})(\xi).$$

We can associate with this system described through state-space representation two input-output maps.

DEFINITION 2.2. Let us consider a system (F, H) defined by the set-valued map F describing the dynamics of the differential inclusion and the observation map H.

Let $\mathcal{S} := \mathcal{S}_F$ denote the solution map of the differential inclusion. We shall say that

1) The product $I_- := H \circ \mathcal{S}$, from X to $\mathscr{C}(0, T; Y)$ defined by

$$\forall x_0 \in X, \quad I_-(x_0) := \bigcup_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot))$$

is the **sharp** input-output map.

2) The "square product" $I_+ := H \square S$, from X to $\mathscr{C}(0, T; Y)$ defined by

$$\forall x_0 \in X, \quad I_+(x_0) := \bigcap_{x(\cdot) \in \mathcal{S}(x_0)} H(x(\cdot))$$

is the **hazy** input-output map.

Remark. Observe that when the observation map is single-valued, the use of a nontrivial hazy input-output map requires that all solutions $x(\cdot) \in \mathcal{S}(x_0)$ yield the

same observation $y(\cdot) = H(x(\cdot))$. Hence we have to study when this possibility occurs, by projecting the differential inclusion (15) onto a differential equation that "tracks" all the solutions to the differential inclusion. This is the purpose of the next section. □

3. Projection of a system onto the observation space. Our first task is to provide conditions implying that the hazy input-output map $I_+ := H \square \mathcal{S}$ is not trivial, above all when the observation map is single-valued.

We shall tackle this issue by "projecting" the differential inclusion given in the state space X onto a differential inclusion in the observation space Y in such a way that solutions to the projected differential inclusion are observations of solutions to the original differential inclusion.

Let us consider a differential inclusion

$$x'(t) \in F(t, x(t)), \qquad x(0) = x_0 \tag{16}$$

where $F: [0, T] \times X \rightsquigarrow X$ is a nontrivial set-valued map and an observation map $H: X \rightsquigarrow Y$ from X to another finite-dimensional vector-space Y.

We project the differential inclusion (16) to a differential inclusion (or a differential equation) in the observation space Y described by a set-valued map G (or a single-valued map g)

$$y'(t) \in G(t, y(t)) \quad (\text{or } y'(t) = g(t, y(t))), \quad y(0) = y_0 \tag{17}$$

that allows us to track partially or completely solutions $x(\cdot)$ to the differential inclusion (16) in the following sense:

(18)
- (a) $\forall (x_0, y_0) \in \text{Graph}(H)$ there exist solutions $x(\cdot)$ and $y(\cdot)$ to (16) and (17) such that $\forall t \in [0, T]$, $y(t) \in H(x(t))$,
- (b) $\forall (x_0, y_0) \in \text{Graph}(H)$ all solutions $x(\cdot)$ and $y(\cdot)$ to (16) and (17) satisfy $\forall t \in [0, T]$, $y(t) \in H(x(t))$.

The second property means that the differential inclusion (17) is, so to speak, "blind" to the solutions to the differential inclusion (16). When it is satisfied, we see that for all $x_0 \in H^{-1}(y_0)$, all the solutions to the differential inclusion (16) do satisfy

$$\forall t \in [0, T], \quad y(t) \in H(x(t)).$$

We need the following definition.

DEFINITION 3.1. Let (x, y) belong to the graph of a set-valued map $F: X \rightsquigarrow Y$ from a normed space X to another Y. Then the **contingent derivative** $DF(x, y)$ of F at (x, y) is the set-valued map from X to Y defined by

$$v \in DF(x, y)(u) \Leftrightarrow \liminf_{h \to 0+, u' \to u} d\left(v, \frac{F(x + hu') - y}{h}\right) = 0$$

and the **paratingent derivative** $PF(x, y)$ of F at (x, y) is the set-valued map from X to Y defined by

$$v \in PF(x, y)(u) \Leftrightarrow \liminf_{h \to 0+, (x', y') \to_F (x, y), u' \to u} d\left(v, \frac{F(x' + hu') - y'}{h}\right) = 0$$

where $\to_F$ denotes the convergence in Graph (F).

(See [23] for the study of paratingent cones and the applications of Choquet's Theorem.)

When F is pseudo-Lipschitzian around $(x, y) \in \text{Graph}(F)$ in the sense that

$$\exists k > 0 \text{ such that } \forall (x', y') \in \text{Graph } F \text{ near } (x, y), \forall x'' \in X \text{ near } x,$$
$$F(x'') \subset F(x') + k\|x' - x''\| B$$

the above formulas become

$$\text{(i)}\quad v \in DF(x,y)(u) \Leftrightarrow \liminf_{h\to 0+} d\left(v, \frac{F(x+hu)-y}{h}\right) = 0,$$

$$\text{(ii)}\quad v \in PF(x,y)(u) \Leftrightarrow \liminf_{h\to 0+,(x',y')\to_F(x,y)} d\left(v, \frac{F(x'+hu)-y'}{h}\right) = 0.$$

Moreover, in this case the derivative $DF(x, y)$ has nonempty images and is k-Lipschitzian (see [12]).

PROPOSITION 3.1. *Let us consider a closed set-valued map H from X to Y.*

1. *Let us assume that F and G are nontrivial upper semicontinuous set-valued maps with nonempty compact convex images and with linear growth. We posit the assumption*

$$\forall (x,y) \in \text{Graph}\,(H),\quad G(t,y) \cap (DH(x,y) \circ F)(t,x) \neq \emptyset. \tag{19}$$

Then property (18)(a) *holds true.*

2. *Let us assume that $F \times G$ is nontrivial Lipschitzian on a neighborhood of* $[0, T] \times$ Graph (H) *and has a linear growth. We posit the assumption*

$$\forall (x,y) \in \text{Graph}\,(H),\quad G(t,y) \subset (DH(x,y) \square F)(t,x). \tag{20}$$

Then property (18)(b) *is satisfied.*

Proof. It follows obviously from the viability and invariance theorems of the graph of H for the set-valued map $F \times G$.

1. When $G(t, y)$ intersects $(DH(x, y) \circ F)(t, x) = \cup_{v \in F(t,x)} DH(x, y)(v)$, we deduce that Graph (H) is a viability domain of $F \times G$ $(t, \cdot)$. Hence we apply the Viability Theorem (see [14], [1, Thm. 4.2.1, p. 180]).

2. When $F \times G$ is Lipschitzian and satisfies (20), we deduce that Graph (H) is invariant by $F \times G(t, \cdot)$. Hence we apply the Invariance Theorem (see [8], [1, Thm. 4.6.2]). □

In particular, we have obtained a sufficient condition for the hazy input-output set-valued map I_+ to be nontrivial.

First, it will be convenient to introduce the following definition.

DEFINITION 3.2. Let us consider $F : [0, T] \times X \rightsquigarrow X$ and $H : [0, T] \times X \rightsquigarrow Y$. We shall say that a nontrivial set-valued map $G : [0, T] \times Y \rightsquigarrow Y$ is a **Lipschitzian square projection** of a set-valued map $F : [0, T] \times X \rightsquigarrow X$ by H if and only if

(i) $F \times G$ is Lipschitzian around $[0, T] \times$ Graph (H),

(ii) $\forall (x, y) \in \text{Graph}\,(H)$, $G(t, y) \subset (DH(x, y) \square F)(t, x)$.

Therefore, for being able to use nontrivial hazy input-output maps, we shall use the following consequence of Proposition 3.1.

PROPOSITION 3.2. *Let us assume that $F : [0, T] \times X \rightsquigarrow X$ and $H : X \rightsquigarrow Y$ are given. If there exists a Lipschitzian square projection of F by H, then the hazy input-output map $I_+ := H \square \mathcal{S}$ has nonempty values for any initial value $y_0 \in H(x_0)$.*

Remark. When the observation map H is single-valued and differentiable, then conditions (19) and (20) become, respectively,

(i) $\forall y \in H^{-1}(x)$, $G(t, y) \cap (\cup_{v \in F(t,x)} H'(x)(v)) \neq \emptyset$
or $G(t, y) \cap (H'(x) \circ F)(t, x) \neq \emptyset$,

(ii) $\forall y \in H^{-1}(x)$, $G(t, y) \subset \cap_{v \in F(t,x)} H'(x)(v)$
$=: (H'(x) \square F)(t, x)$.

When $G = g$ is a single-valued map, we obtain naturally the following consequence.

COROLLARY 3.1. *Let us consider a closed set-valued map H from X to Y.*

1) *Let us assume that F is an upper semicontinuous set-valued map with nonempty compact convex images and with linear growth and that there exists a continuous selection g with linear growth of the product*

$$\forall (x, y) \in \text{Graph}\,(H), \quad g(t, y) \in (DH(x, y) \circ F)(t, x).$$

Then property (18)(a) *holds true.*

2) *Let us assume that $F \times g$ is Lipschitzian on a neighborhood of* $[0, T] \times \text{Graph}\,(H)$ *with linear growth. If g satisfies*

$$\forall (x, y) \in \text{Graph}\,(H), \quad g(t, y) \in (DH(x, y) \square F)(t, x),$$

then property (18)(b) *is satisfied.*

Remark. Naturally, these formulas have their analogues when the observation maps are time-dependent.

Conditions (19) and (20) becomes, respectively,

(i) $\forall (t, x, y) \in \text{Graph}\,(H)$, $G(t, y) \cap (\cup_{v \in F(t,x)} DH(t, x, y)(1, v)) \neq \emptyset$,

(ii) $\forall (t, x, y) \in \text{Graph}\,(H)$, $G(t, y) \subset \cap_{v \in F(t,x)} DH(t, x, y)(1, v)$.

When the observation map H is single-valued and differentiable, then these conditions can be written in the form

(i) $\forall (t, x) \in \text{Dom}\,(H)$,
$G(t, H(x)) \cap (\partial / \partial t\, H(t, x) + \cup_{v \in F(t,x)} H'_x(t, x)v) \neq \emptyset$, or
$G(t, H(x)) \cap (\partial / \partial t\, H(t, x) + (H'_x(t, x) \circ F)(t, x)) \neq \emptyset$;

(ii) $\forall (t, x) \in \text{Dom}\,(H)$,
$G(t, H(x)) \subset \partial / \partial t\, H(t, x) + \cap_{v \in F(t,x)} H'_x(t, x)v$
$=: \partial / \partial t\, H(t, x) + (H'_x(t, x) \square F)(t, x)$.

Remark. We observe that when the set-valued maps F and G are time-independent, Proposition 3.1 can be reformulated in terms of commutativity of schemes for square products.

Let us denote by Φ the solution map associating to any y_0 a solution to the differential inclusion (equation) (17) starting at y_0 (when G is single-valued and Lipschitzian such solution is unique).

Then we can deduce that property (18)(b) is equivalent to

$$\forall y_0 \in \text{Im}\,(H), \quad \Phi(y_0) \subset ((H \square \mathcal{S}) \square H^{-1})(y_0).$$

Condition (20) becomes: for all $y \in \text{Im}\,(H)$,

$$G(y) \subset \bigcap_{x \in H^{-1}(y)} \bigcap_{v \in F(x)} DH(x, y)(v) := (DH(x, y) \square F) \square H^{-1}(y).$$

In other words, the second part of Proposition 3.1 implies that if the scheme

$$\begin{array}{ccc} & X & \overset{F}{\rightsquigarrow} & X & \\ H^{-1} & \uparrow & \overset{G}{\rightsquigarrow} & \downarrow & DH(x, y) \\ & Y & & Y & \end{array}$$

is "commutative for the square products," then the derived scheme

$$\begin{array}{ccccc} & X & & \overset{\mathscr{S}}{\rightsquigarrow} & \mathscr{C}(0, T; X) \\ H & \Updownarrow & H^{-1} & \downarrow & H \\ & Y & & \overset{\Phi}{\rightsquigarrow} & \mathscr{C}(0, T; Y) \end{array}$$

is also commutative for the square products.

4. Hazy and sharp observability. The observability concepts deal with the possibility of recovering the input—here, the initial state—from the observation of the evolution of the state. In other words, they are related to the injectivity of the sharp and hazy input-output set-valued maps, or, more generally, to the single-valuedness of the inverses of those input-output maps.

So, we start with precise definitions.

DEFINITION 4.1. Let $\mathscr{F}: X \rightsquigarrow Y$ be a set-valued map. We shall say that it enjoys **local inverse single-valuedness at** an element (x^*, y^*) of its graph if and only if there exists a neighborhood $N(x^*)$ such that

$$\{x \,|\, y^* \in \mathscr{F}(x)\} \cap N(x^*) = \{x^*\}.$$

If the neighborhood $N(x^*)$ coincides with the domain of $\mathscr{F}$, $\mathscr{F}$ is said to have **(global) inverse single-valuedness** at y^*.

We shall say that it is **locally injective** around x^* if and only if there exists a neighborhood $N(x^*)$ such that, for all $x_1 \neq x_2 \in N(x^*)$, we have $\mathscr{F}(x_1) \cap \mathscr{F}(x_2) = \varnothing$. It is said to be (**globally**) **injective** if we can take for neighborhood $N(x^*)$ the whole domain of $\mathscr{F}$.

With these definitions at hand, we are able to adapt some of the observability concepts to the set-valued case.

DEFINITION 4.2. Assume that the sharp and hazy input-output maps are defined on nonempty open subsets. Let $y^* \in H(\mathscr{S}(x_0))$ be an observation associated with an initial state x_0.

We shall say that the system is **sharply observable at** (respectively, **locally sharply observable at**) (x_0, y^*) if and only if the sharp input-output map I_- enjoys the global inverse single-valuedness (respectively, **local**) at (x_0, y^*).

Hazily observable and locally hazily observable systems are defined in the same way when the sharp input-output map is replaced by the hazy input-output map I_+.

The system is said to be **hazily (locally) observable around** x_0 if the hazy input-output map I_+ is (*locally*) injective around x_0.

Remarks. Several obvious remarks are in order. We observe that the system is sharply locally observable **at** (x_0, y^*) if and only if there exists a neighborhood $N(x_0)$ of x_0 such that

$$\text{If } x(\cdot) \in \mathscr{S}(N(x_0)) \text{ is such that } y^*(\cdot) \in H(x(\cdot)), \text{ then } x(0) = x_0,$$

i.e., sharp observability at (x_0, y^*) means that an observation $y^*(\cdot)$ characterizes the input x_0.

The system is hazily locally observable **at** (x_0, y^*) if and only if there exists a neighborhood $N(x_0)$ of x_0 such that, for all $x_1 \in N(x_0)$,

$$\text{If } \forall x(\cdot) \in \mathscr{S}(x_1),\ y^*(\cdot) \in H(x(\cdot)), \text{ then } x_1 = x_0.$$

It is also clear that sharp local (respectively, global) observability implies hazy local (respectively, global) observability.

We mention that if we consider two systems $\mathcal{F}_1$ and $\mathcal{F}_2$ such that

$$\forall x \in X, \quad \mathcal{F}_1(x) \subset \mathcal{F}_2(x),$$

then

1. If $\mathcal{F}_2$ is sharply locally (respectively globally) observable, so is $\mathcal{F}_1$;
2. If $\mathcal{F}_1$ is hazily locally (respectively globally) observable, so is $\mathcal{F}_2$.

We shall derive local observability and injectivity of a set-valued map $\mathcal{F}: X \rightsquigarrow Y$ from a general principle based on the differential calculus of set-valued maps.

For that purpose, we shall use its contingent and paratingent derivatives $D\mathcal{F}(x^*, y^*)$ and $P\mathcal{F}(x^*, y^*)$, which are closed processes from X to Y (see the previous section for precise definitions).

Since $0 \in D\mathcal{F}(x^*, y^*)(0)$, we observe that to say that the "linearized map" $D\mathcal{F}(x^*, y^*)$ enjoys the inverse single-valuedness at zero amounts to saying that the inverse image $D\mathcal{F}(x^*, y^*)^{-1}(0)$ contains only one element, i.e., that its kernel $\operatorname{Ker} D\mathcal{F}(x^*, y^*)$ is equal to zero, where the kernel is naturally defined by

$$\operatorname{Ker} D\mathcal{F}(x^*, y^*) := D\mathcal{F}(x^*, y^*)^{-1}(0).$$

THEOREM 4.1. *Let $\mathcal{F}$ be a set-valued map from a finite dimensional vector-space X to a Banach space Y and (x^*, y^*) belong to its graph.*

1. *If the kernel of the contingent derivative $D\mathcal{F}(x^*, y^*)$ of $\mathcal{F}$ at (x^*, y^*) is equal to $\{0\}$, then there exists a neighborhood $N(x^*)$ such that*

$$\{x \mid y^* \in \mathcal{F}(x)\} \cap N(x^*) = \{x^*\}. \tag{21}$$

2. *Let us assume that there exists $\gamma > 0$ such that $\mathcal{F}(x^* + \gamma B)$ is relatively compact and that $\mathcal{F}$ has a closed graph (then $\mathcal{F}(x^* + \gamma B)$ is compact). If for all $y \in \mathcal{F}(x^*)$ the kernels of the paratingent derivatives $P\mathcal{F}(x^*, y)$ of $\mathcal{F}$ at (x^*, y) are equal to $\{0\}$, then $\mathcal{F}$ is locally injective around x^*.*

Proof. 1. Assume that the conclusion (21) is false. Then there exists a sequence of elements $x_n \neq x^*$ converging to x^* satisfying

$$\forall n \geqq 0, \quad y^* \in \mathcal{F}(x_n).$$

Let us set $h_n := \|x_n - x^*\|$, which converges to zero, and

$$u_n := (x_n - x^*)/h_n.$$

The elements u_n do belong to the unit sphere, which is compact. Hence a subsequence (again denoted) u_n does converge to some u different from zero. Since the above equation can be written as

$$\forall n \geqq 0, \quad y^* + h_n 0 \in \mathcal{F}(x^* + h_n u_n)$$

we deduce that

$$0 \in D\mathcal{F}(x^*, y^*)(u).$$

Hence we have proved the existence of a nonzero element of the kernel of $D\mathcal{F}(x^*, y^*)$, which is a contradiction.

2. Assume that $\mathcal{F}$ is not locally injective. Then there exists a sequence of elements $x_n^1, x_n^2 \in N(x^*)$, $x_n^1 \neq x_n^2$, converging to x^* and y_n satisfying

$$\forall n \geqq 0, \quad y_n \in \mathcal{F}(x_n^1) \cap \mathcal{F}(x_n^2).$$

Let us set $h_n := \|x_n^1 - x_n^2\|$, which converges to zero, and

$$u^n := (x_n^1 - x_n^2)/h_n.$$

The elements u_n do belong to the unit sphere, which is compact. Hence a subsequence (again denoted) u_n does converge to some u different from zero.

Then for all large n

$$y_n \in \mathcal{F}(x_n^1) \cap \mathcal{F}(x_n^2) := \mathcal{F}(x_n^2 + h_n u_n) \cap \mathcal{F}(x_n^2) \subset \mathcal{F}(x^* + \gamma B).$$

We deduce that a subsequence (again denoted) y_n converges to some $y \in \mathcal{F}(x^*)$ (because Graph $(\mathcal{F})$ is closed).

Since the above equation implies that

$$\forall n \geqq 0, \quad y_n + h_n 0 \in \mathcal{F}(x_n^2 + h_n u_n)$$

we deduce that

$$0 \in P\mathcal{F}(x^*, y)(u).$$

Hence we have proved the existence of a nonzero element of the kernel of $P\mathcal{F}(x^*, y)$, which is a contradiction. □

When $\mathcal{F}$ is convex (i.e., its graph is convex), we have a simple criterion for global observability: Define the algebraic derivative $D_a\mathcal{F}(x, y)$ of $\mathcal{F}$ at (x, y) by

$$v \in D_a\mathcal{F}(x, y)(u) \Leftrightarrow \exists h > 0 \quad \text{such that } y + hv \in \mathcal{F}(x + hu).$$

Then $\overline{\text{Graph } D_a\mathcal{F}(x, y)} = \text{Graph } D\mathcal{F}(x, y)$.

PROPOSITION 4.1. *Let $\mathcal{F}$ be a convex set-valued map from a Banach space X to a Banach space Y and (x^*, y^*) belong to its graph. If the kernel of $D_aF(x^*, y^*)$ is equal to zero, then*

$$x \neq x^* \Rightarrow y^* \notin \mathcal{F}(x).$$

Proof. If not, there exists $x \neq x^*$ such that $y^* \in \mathcal{F}(x)$. We set $u := x - x^*$. Equality

$$y^* + 0 = y^* \in \mathcal{F}(x) = \mathcal{F}(x^* + u)$$

implies that u, which is different from zero, does belong to the kernel of $D_a\mathcal{F}(x^*, y^*)$. □

Therefore, using this result for proving sufficient conditions for sharp and/or hazy observability, we need:

1. To have chain rule formulas for composition and square products of set-valued maps;
2. To characterize the derivatives of the solution map in terms of solutions to the associated variational equations.

The next proposition provides chain rule formulas for square products that are needed for estimating the contingent and paratingent derivatives of the hazy input-output map I_+ in terms of the adjacent and circatangent derivatives of the map G at (x^*, y^*).

Despite the fact that both adjacent and circatangent derivatives can be defined for any set-valued map F, the formulas are simpler when we deal with pseudo-Lipschitzian set-valued maps. Since we use them only in this context in this paper, we provide their definitions in this limited case.

DEFINITION 4.3. Let (x, y) belong to the graph of a set-valued map $F: X \rightsquigarrow Y$ from a normed space X to another Y. Assume that F is pseudo-Lipschitzian around $(x, y) \in \text{Graph } (F)$, then the **adjacent derivative** $D^\flat F(x, y)$ and the **circatangent derivative** $CF(x, y)$ are the set-valued maps from X to Y, respectively, defined by

$$v \in D^\flat F(x, y)(u) \Leftrightarrow \lim_{h \to 0+} \left(v, \frac{F(x + hu) - y}{h} \right) = 0$$

and

$$v \in CF(x, y)(u) \Leftrightarrow \lim_{h \to 0+, (x', y') \to_F (x, y)} d\left(v, \frac{F(x'+hu) - y'}{h}\right) = 0.$$

Several remarks are in order. First, all these derivatives are positively homogeneous and their graphs are closed.

We observe the obvious inclusions

$$CF(x, y)(u) \subset D^\flat F(x, y)(u) \subset DF(x, y)(u) \subset PF(x, y)(u)$$

and that the definitions of contingent and adjacent derivatives on the one hand, the paratingent and circatangent derivatives, on the other, are symmetric. When $F := f$ is single-valued, we set

$$Df(x) := Df(x, f(x)), \quad D^\flat f(x) := D^\flat f(x, f(x)), \quad Cf(x) := Cf(x, f(x)).$$

We see easily that

$$Df(x)(u) \ni f'(x)u \qquad \text{if } f \text{ is Gâteaux differentiable,}$$

$$D^\flat f(x)(u) = f'(x)u \qquad \text{if } f \text{ is Fréchet differentiable,}$$

$$Cf(x)(u) = f'(x)u \qquad \text{if } f \text{ is continuously differentiable.}$$

The choice of these strange limits is dictated by the fact that the graph of each of these derivatives is the corresponding tangent cone to the graph of F at (x, y). (The graphs of the circatangent derivatives are the Clarke tangent cones to the graphs, which are always convex.)

The most familiar instance of set-valued maps is the inverse of a noninjective single-valued map. The **derivative of the inverse of a set-valued map F is the inverse of the derivative:**

$$P(F)^{-1}(y, x) = PF(x, y)^{-1},$$

$$D(F)^{-1}(y, x) = DF(x, y)^{-1},$$

$$D^\flat(F)^{-1}(y, x) = D^\flat F(x, y)^{-1},$$

$$C(F)^{-1}(y, x) = CF(x, y)^{-1},$$

and enjoy a now well investigated calculus.

The circatangent derivatives are closed convex processes, because their graph are closed convex cones, i.e., they are set-valued analogues of the continuous linear operators. We refer to [21] and [2, Chap. 7] for various properties of closed convex processes.

We say that a set-valued map F is **derivable** at $(x, y) \in \text{Graph}(F)$ if $DF(x, y) = D^\flat F(x, y)$ and that it is **derivable** if it is derivable at every point of its graph.

We say that a set-valued map F is **sleek** at $(x, y) \in \text{Graph}(F)$ if

$$\text{Graph}(F) \ni (x', y') \rightsquigarrow \text{Graph}(DF(x', y')) \text{ is lower semicontinuous at } (x, y)$$

and it is **sleek** if it is sleek at every point of its graph. In this case, we can prove that **the contingent, adjacent, and circatangent derivatives coincide.**

PROPOSITION 4.2. *Let us consider a set-valued map G from a Banach space X to a Banach space Y and a single-valued map H from Y to a Banach space Z. Assume that G is Lipschitzian around x^*. If H is differentiable at some $y^* \in G(x^*)$, then*

1. *The contingent derivative of $H \square G$ is contained in the square product of the derivative of H and the adjacent derivative of G:*

$$\forall u \in \mathrm{Dom}\,(D^{\flat}G(x^*, y^*)), \quad D(H \square G)(x^*, H(y^*))(u) \subset H'(y^*) \square D^{\flat}G(x^*, y^*)(u);$$

2. *If H is continuously differentiable around y^*, then the paratingent derivative of $H \square G$ is contained in the square product of the derivative of H and the circatangent derivative of G:*

$$\forall u \in \mathrm{Dom}\,(CG(x^*, y^*)), \quad P(H \square G)(x^*, H(y^*))(u) \subset H'(y^*) \square CG(x^*, y^*)(u).$$

Proof. 1. Let $u \in \mathrm{Dom}\, D^{\flat}G(x^*, y^*)$ and w belong to $D(H \square G)(x^*, H(y^*))(u)$. Hence there exist a sequence $h_n > 0$ converging to zero and sequences of elements u_n, w_n converging to u and w, respectively, such that

$$\forall n \geqq 0, \quad H(y^*) + h_n w_n \in \bigcap_{y \in G(x^* + h_n u_n)} H(y).$$

Take now any v in $D^{\flat}G(x^*, y^*)(u)$. Since G is Lipschitzian around x^*, there exists a sequence of elements v_n converging to v such that

$$\forall n \geqq 0, \quad y^* + h_n v_n \in G(x^* + h_n u_n).$$

Therefore,

$$\forall n \geqq 0, \quad H(y^*) + h_n w_n = H(y^* + h_n v_n).$$

Since H is differentiable at y^*, we infer that

$$H'(y^*)v = w.$$

Since this is true for every element v of $D^{\flat}G(x^*, y^*)(u)$, we deduce that

$$w \in \bigcap_{v \in D^{\flat}G(x^*, y^*)(u)} H'(y^*)v = H'(y^*) \square D^{\flat}G(x^*, y^*)(u).$$

2. Let $u \in \mathrm{Dom}\, CG(x^*, y^*)$ and w belong to $P(H \square G)(x^*, H(y^*))(u)$. Hence there exist a sequence $h_n > 0$ converging to zero and sequences of elements $(x_n, z_n) \in \mathrm{Graph}\,(H \square G)$, u_n and w_n converging to (x^*, z^*), u and w, respectively, such that

$$\forall n \geqq 0, \quad z_n + h_n w_n \in \bigcap_{y \in G(x_n + h_n u_n)} H(y).$$

The set-valued map G being Lipschitzian, there exists a sequence of elements $y_n \in G(x_n)$ converging to y^*. By definition of the square product, we know that $z_n = H(y_n)$.

Now take any v in $CG(x^*, y^*)(u)$. Since G is Lipschitzian around x^*, there exists a sequence of elements v_n converging to v such that

$$\forall n \geqq 0, \quad y_n + h_n v_n \in G(x_n + h_n u_n).$$

Therefore,

$$\forall n \geqq 0, \quad H(y_n) + h_n w_n = H(y_n + h_n v_n).$$

Since H is continuously differentiable around y^*, we infer that

$$H'(y^*)v = w.$$

Since this is true for every element v of $CG(x^*, y^*)(u)$, we deduce that

$$w \in \bigcap_{v \in CG(x^*, y^*)(u)} H'(y^*)v = H'(y^*) \square CG(x^*, y^*)(u). \qquad \square$$

For the usual product, it is easy to check that

$$H'(y) \circ DG(x, y)(u) \subset D(H \circ G)(x, H(y))(u).$$

Naturally, equality holds true for algebraic derivatives: If $H \in \mathscr{L}(Y, Z)$ is a linear operator, we check that

$$H \circ D_a G(x, y)(u) = D_a(H \circ G)(x, H(y))(u). \tag{22}$$

We do not know for the time other elegant criteria implying the chain rule for the usual composition product of set-valued maps in infinite-dimensional spaces. Let us mention, however, the following result involving the **co-subdifferential** $DG(x_0, y_0)^{0*}$, which is the closed convex process from Y^* to X^* defined by

$$p \in DG(x, y)^{0*}(q) \quad \text{if and only if } \forall (x', y') \in \text{Graph}\,(G), \quad \langle p, x' - x\rangle \leqq \langle q, y' - y\rangle.$$

Let us assume that H is a continuous linear operator $H \in \mathscr{L}(Y, Z)$ from Y to Z. Equality

$$D(H \circ G)(x_0, Hy_0)(u) = \overline{H \circ DG(x_0, y_0)(u)}$$

holds true if X and Y are reflexive Banach spaces and the co-subdifferential of G at (x_0, y_0) satisfies

$$\text{Im}\,(H^*) + \text{Dom}\,(DG(x_0, y_0)^{0*}) = Y^*.$$

Furthermore, this condition implies that the kernels of $D(H \circ G)(x_0, Hy_0)$ and $\overline{H \circ DG(x_0, y_0)}$ are equal to $\{0\}$ (see [6]).

5. Variational inclusions. We now provide estimates of the contingent, adjacent, and circatangent derivatives of the solution map $\mathscr{S}$ associated to the differential inclusion

$$x'(t) \in F(t, x(t)), \quad \text{a.e. in } [0, T]. \tag{23}$$

We shall express these estimates in terms of the solution maps of adequate linearizations of differential inclusion (23) of the form

$$w'(t) \in F'(t, x(t), x'(t))(w(t))$$

where for almost all t, $F'(t, x, y)(u)$ denotes one of the (contingent, adjacent, or circatangent) derivatives of the set-valued map $F(t, \cdot, \cdot)$ at a point (x, y) of its graph (in this section the set-valued map F is regarded as a family of set-valued maps $x \rightsquigarrow F(t, x)$ and *the derivatives are taken with respect to the state variable only*).

These linearized differential inclusions can be called the **variational inclusions**, since they extend (in various ways) the classical variational equations of ordinary differential equations.

Let $\bar{x}$ be a solution of the differential inclusion (23). We assume that F satisfies the following assumptions:

$$\begin{aligned} &\text{(i)} && \forall x \in X \text{ the set-valued map } F(\cdot, x) \text{ is measurable,} \\ &\text{(ii)} && \forall t \in [0, T],\ \forall x \in X,\ F(t, x) \text{ is a nonempty closed set,} \\ &\text{(iii)} && \exists \beta > 0,\ k(\cdot) \in L^1(0, T) \text{ such that for almost all } t \in [0, T] \\ & && \text{the map } F(t, \cdot) \text{ is } k(t) - \text{Lipschitz on } \bar{x}(t) + \beta B. \end{aligned} \tag{24}$$

Under the above assumptions the map $x_0 \rightsquigarrow \mathscr{S}(x_0)$ is pseudo-Lipschitzian around $(\bar{x}(0), \bar{x})$. Consider the **adjacent variational inclusion**, which is the "linearized" along the trajectory $\bar{x}$ inclusion

$$\begin{aligned} &w'(t) \in D^\flat F(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. in } [0, T], \\ &w(0) = u \end{aligned} \tag{25}$$

where $u \in X$. In Theorems 5.1 and 5.2 below we consider the solution map $\mathscr{S}$ as the set-valued map from $\mathbf{R}^n$ to the Sobolev space $W^{1,1}(0, T; \mathbf{R}^n)$.

THEOREM 5.1 (adjacent variational inclusion [11]). *If the assumptions* (24) *hold true, then for all* $u \in X$, *every solution* $w \in W^{1,1}(0, T; X)$ *to the linearized inclusion* (25) *satisfies* $w \in D^\flat \mathcal{S}(\bar{x}(0), \bar{x})(u)$.

In other words,

$$\{w(\cdot) \mid w'(t) \in D^\flat F(t, \bar{x}(t), \bar{x}'(t))(w(t)),\ w(0) = u\} \subset D^\flat \mathcal{S}(\bar{x}(0), \bar{x})(u).$$

Proof. Filippov's theorem (see, for example, [1, Thm. 2.4.1, p. 120]) implies that the map $u \to \mathcal{S}(u)$ is pseudo-Lipschitzian on a neighborhood of $(\bar{x}(0), \bar{x})$. Let $h_n > 0$, $n = 1, 2, \ldots$ be a sequence converging to zero. Then, by the very definition of the adjacent derivative, for almost all $t \in [0, T]$,

$$\lim_{n\to\infty} d\left(w'(t), \frac{F(t, \bar{x}(t) + h_n w(t)) - \bar{x}'(t)}{h_n}\right) = 0. \tag{26}$$

Moreover, since $\bar{x}'(t) \in F(t, \bar{x}(t))$ almost everywhere in $[0, T]$, by (24), for all sufficiently large n and almost all $t \in [0, T]$

$$d(\bar{x}'(t) + h_n w'(t), F(t, \bar{x}(t) + h_n w(t))) \leqq h_n(\|w'(t)\| + k(t)\|w(t)\|).$$

This, (26), and the Lebesgue dominated convergence theorem yield

$$\int_0^T d(\bar{x}'(t) + h_n w'(t), F(t, \bar{x}(t) + h_n w(t)))\, dt = o(h_n) \tag{27}$$

where $\lim_{n\to\infty} o(h_n)/h_n = 0$. By the Filippov Theorem (see, for example, [1, Thm. 2.4.1, p. 120]) and by (27) there exist $M \geqq 0$ and solutions $y_n \in \mathcal{S}(\bar{x}(0) + h_n u)$ satisfying

$$\|y_n' - \bar{x}' - h_n w'\|_{L^1(0,T;X)} \leqq M o(h_n).$$

Since $(y_n(0) - \bar{x}(0))/h_n = u = w(0)$ this implies that

$$\lim_{n\to\infty} \frac{y_n - \bar{x}}{h_n} = w \quad \text{in } C(0, T; X), \qquad \lim_{n\to\infty} \frac{y_n' - \bar{x}'}{h_n} = w' \quad \text{in } L^1(0, T; X).$$

Hence

$$\lim_{n\to\infty} d\left(w, \frac{\mathcal{S}(\bar{x}(0) + h_n u) - \bar{x}}{h_n}\right) = 0.$$

Since u and w are arbitrary the proof is complete. □

Next consider the **circatangent variational inclusion,** which is the linearization involving circatangent derivatives:

$$\begin{aligned} &w'(t) \in CF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. in } [0, T], \\ &w(0) = u \end{aligned} \tag{28}$$

where $u \in X$. The next theorem is similar to Lemma 4.11 of [11].

THEOREM 5.2 (circatangent variational inclusion). *Assume that conditions* (24) *hold true. Then for all* $u \in X$, *every solution* $w \in W^{1,1}(0, T; X)$ *to the linearized inclusion* (28) *satisfies* $w \in C\mathcal{S}(\bar{x}(0), \bar{x})(u)$.

In other words,

$$\{w(\cdot) \mid w'(t) \in CF(t, \bar{x}(t), \bar{x}'(t))(w(t)),\ w(0) = u\} \subset C\mathcal{S}(\bar{x}(0), \bar{x})(u).$$

Proof. By Filippov's theorem the map $u \to \mathcal{S}(u)$ is pseudo-Lipschitzian on a neighborhood of $(\bar{x}(0), \bar{x})$. Consider a sequence x_n of trajectories of (23) converging to $\bar{x}$ in $W^{1,1}(0, T; X)$ and let $h_n \to 0+$. Then there exists a subsequence $x_j := x_{n_j}$ such that

$$\lim_{j\to\infty} x_j'(t) = \bar{x}'(t) \quad \text{a.e. in } [0, T]. \tag{29}$$

Set $\lambda_j = h_{n_j}$. Then, by definition of circatangent derivative and by (29), for almost all $t \in [0, T]$

$$\lim_{j\to\infty} d\left(w'(t), \frac{F(t, x_j(t)+\lambda_j w(t)) - x_j'(t)}{\lambda_j}\right) = 0. \tag{30}$$

Moreover, using the fact that $x_j'(t) \in F(t, x_j(t))$ almost everywhere in $[0, T]$, we obtain that for almost all $t \in [0, T]$ and for all j large enough

$$d\left(x_j'(t)+\lambda_j w'(t), F(t, x_j(t)+\lambda_j w(t))\right) \leqq \lambda_j(\|w'(t)\| + k(t)\|w(t)\|).$$

This, (30), and the Lebesgue dominated convergence theorem yield

$$\int_0^T d\left(x_j'(t)+\lambda_j w'(t), F(t, x_j(t)+\lambda_j w(t))\right) dt = o(\lambda_j) \tag{31}$$

where $\lim_{j\to\infty} o(\lambda_j)/\lambda_j = 0$. By the Filippov Theorem and (31), there exist $M \geqq 0$ and solutions $y_j \in \mathcal{S}(x_j(0)+\lambda_j u)$ satisfying

$$\|y_j' - x_j' - \lambda_j w'\| L^1(0, T; X) \leqq M o(\lambda_j).$$

Since $(y_j(0) - x_j(0))/\lambda_j = u = w(0)$, this implies that

$$\lim_{j\to\infty} \frac{y_j - x_j}{h_{n_j}} = w \quad \text{in } C(0, T; X), \qquad \lim_{j\to\infty} \frac{y_j' - x_j'}{h_{n_j}} = w' \quad \text{in } L^1(0, T; X).$$

Hence

$$\lim_{j\to\infty} d\left(w, \frac{\mathcal{S}(x_j(0)+h_{n_j}u) - x_j}{h_{n_j}}\right) = 0. \tag{32}$$

Therefore we have proved that for every sequence of solutions x_n to (23) converging to $\bar{x}$ and every sequence $h_n \to 0+$, there exists a subsequence $x_j = x_{n_j}$ that satisfies (32). This yields that for every sequence of solutions x_n converging to $\bar{x}$ and $h_n \to 0+$

$$\lim_{n\to\infty} d\left(w, \frac{\mathcal{S}(x_n(0)+h_n u) - x_n}{h_n}\right) = 0.$$

Since u and w are arbitrary the proof is complete. □

We now consider the **contingent variational inclusion**

$$\begin{aligned} &w'(t) \in \overline{\text{co}}\, DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. in } [0, T], \\ &w(0) = u. \end{aligned} \tag{33}$$

THEOREM 5.3. (contingent variational inclusion). *Let us consider the solution map $\mathcal{S}$ as a set-valued map from $\mathbf{R}^n$ to $W^{1,\infty}(0, T; \mathbf{R}^n)$ supplied with the weak-$*$ topology and let $\bar{x}(\cdot)$ be a solution of the differential inclusion* (23) *starting at x_0. Then the contingent derivative $D\mathcal{S}(x_0, \bar{x}(\cdot))$ of the solution map is contained in the solution map of the contingent variational inclusion* (33), *in the sense that*

$$D\mathcal{S}(x_0, \bar{x}(\cdot))(u) \subset \{w(\cdot) \mid w'(t) \in \overline{\text{co}}\, DF(t, \bar{x}(t), \bar{x}'(t))(w(t)),\ w(0) = u\}. \tag{34}$$

To prove the above theorem we need to recall a property of Kuratowski's upper limit: Let K_n be a sequence of subsets of a Banach space X. We say that the set

$$\operatorname*{co\text{-}lim\,sup}_{n\to\infty} K_n := \bigcap_{N>0} \overline{\text{co}} \bigcup_{n>N} K_n$$

is the **convex upper limit** of the sequence K_n. Recall that the **Kuratowski upper limit** of the K_n's is defined by

$$\limsup_{n\to\infty} K_n := \bigcap_{\varepsilon>0} \bigcap_{N>0} \bigcup_{n\geqq N} (K_n + \varepsilon B).$$

It is clear that the convex upper limit is closed and convex. Moreover, since $\operatorname{co} \bigcup_{n\geqq N} (K_n + \varepsilon B) = \operatorname{co} \bigcup_{n\geqq N} K_n + \varepsilon B$, we obtain

$$\operatorname{co\text{-}lim\,sup}_{n\to\infty} K_n := \bigcap_{\varepsilon>0} \bigcap_{N>0} \overline{\operatorname{co}} \bigcup_{n>N} (K_n + \varepsilon B).$$

Hence the convex upper limit contains the closed convex hull of the Kuratowski upper limit. The convex hull of an upper limit and the convex upper limit are related by the following lemma.

LEMMA 5.1. *Let us consider a sequence of subsets K_n contained in a bounded subset of a finite-dimensional vector-space X. Then*

$$\operatorname{co\text{-}lim\,sup}_{n\to\infty} K_n = \overline{\operatorname{co}}(\limsup_{n\to\infty} K_n).$$

Proof. Since an element x of co-lim $\sup_{n\to\infty} K_n$ is the limit of a subsequence of convex combinations v_N of elements of $\bigcup_{n>N} K_n$ and since the dimension of X is an integer p, Carathéodory's Theorem allows us to write

$$v_N := \sum_{j=0}^{p} a_j^N x_{N_j} \quad \text{where } \sum_{j=0}^{p} a_j^N = 1, \quad a_j^N \geqq 0$$

where $N_j \geqq N$ and where x_{N_j} belongs to K_{N_j}. The vector a^N of $p+1$ components a_j^N contains a converging subsequence (again denoted) a^N that converges to some non-negative vector a of $p+1$ components a_j such that $\sum_{j=0}^{p} a_j = 1$.

The subsets K_n being contained in a compact subset, we can extract successively subsequences (again denoted) x_{N_j} converging to elements x_j that belong to the Kuratowski upper limit of the subsets K_n. Hence x is equal to the convex combination $\sum_{j=0}^{p} a_j x_j$ and the lemma is proved. □

Proof. Fix a direction $u \in \mathbf{R}^n$ and let $w(\cdot)$ belong to $D\mathscr{S}(x_0, \bar{x}(\cdot))(u)$. By definition of the contingent derivative, there exist sequences of elements $h_n \to 0+$, $u_n \to u$ and $w_n(\cdot) \to w(\cdot)$ in the weak-* topology of $W^{1,\infty}(0, T; \mathbf{R}^n)$ and $c > 0$ satisfying

$$\begin{aligned} &\text{(i)} \quad \|w_n'(t)\| \leqq c \quad \text{a.e. in } [0, T], \\ &\text{(ii)} \quad \bar{x}'(t) + h_n w_n'(t) \in F(t, \bar{x}(t) + h_n w_n(t)) \quad \text{a.e. in } [0, T], \\ &\text{(iii)} \quad w_n(0) = u_n. \end{aligned} \tag{35}$$

Hence

$$\begin{aligned} &\text{(i)} \quad w_n(\cdot) \text{ converges pointwise to } w(\cdot), \\ &\text{(ii)} \quad w_n'(\cdot) \text{ converges weakly in } L^1(0, T; \mathbf{R}^n) \text{ to } w'(\cdot). \end{aligned} \tag{36}$$

By Mazur's Theorem and (36)(ii), a sequence of convex combinations

$$v_m(t) := \sum_{p=m}^{\infty} a_m^p w_p'(t)$$

converges strongly to $w'(\cdot)$ in $L^1(0, T; X)$. Therefore a subsequence (again denoted) $v_m(\cdot)$ converges to $w'(\cdot)$ almost everywhere. By (35)(i),(ii) for all p and almost all $t \in [0, T]$

$$w_p'(t) \in \frac{F(t, \bar{x}(t) + h_p w_p(t)) - \bar{x}'(t)}{h_p} \cap cB.$$

Let $t \in [0, T]$ be a point where $v_m(t)$ converges to $w'(t)$ and $x'(t) \in F(t, x(t))$. Fix an integer $n \geqq 1$ and $\varepsilon > 0$. By (36)(i), there exists m such that $h_p \leqq 1/n$ and $\|w_p(t) - w(t)\| \leqq 1/n$ for all $p \geqq m$.

Then, by setting

$$\Phi(y, h) := \frac{F(t, \bar{x}(t) + hy) - \bar{x}'(t)}{h} \cap cB$$

we obtain that

$$v_m(t) \in K_n := \operatorname{co}\left(\bigcup_{h \in]0, 1/n],\ y \in w(t) + B/n} \Phi(y, h) \right)$$

and therefore, by letting m go to ∞, that

$$w'(t) \in \overline{\operatorname{co}}\left(\bigcup_{h \in]0, 1/n],\ y \in w(t) + B/n} \Phi(y, h) \right).$$

Since this is true for any n, we deduce that $w'(t)$ belongs to the convex upper limit:

$$w'(t) \in \bigcap_{n \geqq 1} \overline{\operatorname{co}}\left(\bigcup_{h \in]0, 1/n],\ y \in w(t) + B/n} \Phi(y, h) \right).$$

Since the subsets $\Phi(y, h)$ are contained in the ball of radius c, we infer from the last lemma that $w'(t)$ belongs to the closed convex hull of the Kuratowski upper limit

$$w'(t) \in \overline{\operatorname{co}} \bigcap_{\varepsilon > 0, n \geqq 1} \left(\bigcup_{h \in]0, 1/n],\ y \in w(t) + B/n} (\Phi(y, h) + \varepsilon B) \right).$$

We observe now that

$$\bigcap_{\varepsilon > 0, n \geqq 1} \left(\bigcup_{h \in]0, 1/n],\ y \in w(t) + B/n} (\Phi(y, h) + \varepsilon B) \right) \subset DF(t, \bar{x}(t), \bar{x}'(t))(w(t))$$

to conclude that $w(\cdot)$ is a solution to the differential inclusion

$$w'(t) \in \overline{\operatorname{co}}\, DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. in } [0, T],$$
$$w(0) = u.$$

Since $w \in D\mathcal{S}(x_0, \bar{x}(\cdot))(u)$ is arbitrary we proved (34).

6. Local observability theorems. In this section we piece together the general principle on local inverse univocity and local injectivity (Theorem 4.1), the chain rule formulas (Proposition 4.2) and the estimates of the derivatives of the solution map in terms of solution maps of the variational equations (Theorems 5.1, 5.2 and 5.3) to prove the statements we have announced on local hazy and sharp observability.

Throughout the whole section we assume that H is differentiable and F has a linear growth. We impose also some regularity assumptions on the derivatives of F. In the next theorem it is assumed that $F(t, \cdot)$ is derivable in the sense that its contingent and adjacent derivatives do coincide.

THEOREM 6.1. *Let us assume that for every $t \in [0, T]$, $F(t, \cdot)$ is derivable, satisfies assumptions* (12), *and that it has a Lipschitzian square projection G by H. Let $\bar{x}(\cdot) \in \mathcal{S}(x_0)$. If the contingent variational inclusion*

(37) *for almost all $t \in [0, T]$, $w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t))$*

is globally hazily observable through $H'(\bar{x}(\cdot))$ at zero, then the system (23) *is locally hazily observable through H* **at** *$(x_0, H(\bar{x}))$.*

Proof. We apply the general principle (Theorem 4.1) to the hazy input-output map $I_+ := H \square \mathcal{S}$, which is defined since we have assumed that there exists a square projection G (see Definition 3.2 and Proposition 3.2). We have to prove that the kernel of the contingent derivative $DI_+(x_0, y_0)$ of I_+ (where $y_0 := H(\bar{x}(\cdot))$) is equal to zero. By Filippov's Theorem, the solution map $\mathcal{S}$ is Lipschitzian around x_0. Then we can apply Proposition 4.2 which states that for all $u \in \mathrm{Dom}\,(D^\flat \mathcal{S}(x_0, \bar{x}(\cdot)))$

$$DI_+(x_0, y_0)(u) \subset (H'(\bar{x}(\cdot)) \square D^\flat \mathcal{S}(x_0, \bar{x}(\cdot)))(u).$$

By Theorem 5.1, we know that for any $u \in X$, the set $\Phi(u)$ of solutions to the adjacent variational inclusion (25) starting at u is contained in the adjacent derivative of $\mathcal{S}$:

$$\begin{aligned} \Phi(u) &:= \{w(\cdot) \mid w'(t) \dot{\in} D^\flat F(t, \bar{x}(t), \bar{x}'(t))(w(t)) \text{ and } w(0) = u\} \\ &= \{w(\cdot) \mid w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \text{ and } w(0) = u\} \qquad (38) \\ &\subset D^\flat \mathcal{S}(x_0, \bar{x})(u). \end{aligned}$$

We also know that for all $(x, y) \in \mathrm{Graph}\,(F(t, \cdot))$, the contingent derivative $DF(tx, y)$ is $k(t)$-Lipschitz (see [12]). Hence, by the Filippov Theorem [1, Thm. 2.4.1, p. 120] for every $u \in \mathbf{R}^n$, the contingent variational inclusion (37) has a solution starting at u. Therefore, by (38), $\mathrm{Dom}\,(D^\flat \mathcal{S}(x_0, \bar{x}(\cdot)))$ is equal to the whole space. This yields

$$\forall u \in \mathbf{R}^n, \quad DI_+(x_0, y_0)(u) \subset (H'(\bar{x}) \square \Phi)(u)$$

so that the kernel of $DI_+(x_0, y_0)$ is contained in the kernel of $H'(\bar{x}) \square \Phi$. But to say that the kernel of $H'(\bar{x}) \square \Phi$ is equal to zero amounts to saying that the linearized system (37) is hazily globally observable at zero through $H'(\bar{x}(\cdot))$. Hence the kernel of $DI_+(x_0, y_0)$ is equal to zero, and thus, the inverse image of hazy input-output map $I_+^{-1}(y_0)$ contains locally a unique element x_0. □

Remark. The above result remains true with DF in (37) replaced by $D^\flat F$ if instead of derivability of F we assume that

$$\mathrm{Dom}\,(D^\flat \mathcal{S}(x_0, \bar{x}(\cdot))) = \mathbf{R}^n.$$

In the next theorem we assume that F is sleek, so that its contingent and circatangent derivatives do coincide.

THEOREM 6.2. *Let us assume that F is sleek, has convex images, satisfies assumptions* (12), *and that it has a Lipschitzian square projection G by H. If for all $\bar{x}(\cdot) \in \mathcal{S}(x_0)$ the contingent variational inclusion* (37) *is globally hazily observable through $H'(\bar{x}(\cdot))$* **at** 0, *then the system* (23) *is hazily observable through H* **around** x_0.

Proof. We apply the second part of the general principle on local injectivity (Theorem 4.1) to the hazy input-output map $I_+ := H \square \mathcal{S}$, which is defined since we have assumed that there exists a square projection G. We have to prove that the kernels of the paratingent derivatives $PI_+(x_0, y)$ of I_+ are equal to zero (where $y(\cdot) := H(\bar{x}(\cdot))$ and $\bar{x}(\cdot) \in \mathcal{S}(x_0)$). In the way similar to Theorem 2.2.1 of [1, p. 104], we prove that for all $\gamma > 0$, the set $\mathcal{S}(x_0 + \gamma B)$ is compact in $C(0, T; \mathbf{R}^n)$. Hence $I_+(x_0 + \gamma B)$ is relatively compact in $C(0, T; \mathbf{R}^n)$. By Filippov's Theorem, the solution map $\mathcal{S}$ is Lipschitzian around x_0. This and compactness of $\mathcal{S}(x_0 + \gamma B)$ imply that Graph (I_+) is a closed set. Then we can apply the second part of Proposition 4.2 which states that for all $u \in \mathrm{Dom}\,(C\mathcal{S}(x_0, \bar{x}(\cdot)))$

$$PI_+(x_0, y)(u) \subset H'(\bar{x}(\cdot)) \square C\mathcal{S}(x_0, \bar{x}(\cdot))(u).$$

By Theorem 5.2, we know that for all u, the set $\Phi(u)$ of solutions to the circatangent variational inclusion (24) starting at u is contained in the circatangent derivative of $\mathcal{S}$:

$$\begin{aligned}\Phi(u) &:= \{w(\cdot) \mid w'(t) \in CF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \text{ and } w(0) = u\} \\ &= \{w(\cdot) \mid w'(t) \in DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \text{ and } w(0) = u\} \\ &\subset C\mathcal{S}(x_0, \bar{x})(u).\end{aligned}$$

But from the proof of Theorem 6.1 we know that $\mathrm{Dom}\,(\Phi) = \mathbf{R}^n$. Therefore,

$$PI_+(x_0, y)(u) \subset (H'(\bar{x}) \square \Phi)(u)$$

so that the kernel of $PI_+(x_0, y)$ is contained in the kernel of $H'(\bar{x}) \square \Phi$. But to say that the kernel of $H'(\bar{x}) \square \Phi$ is equal to zero amounts to saying that the linearized system (37) is globally hazily observable through $H'(\bar{x})$ at zero. Hence the kernel of $PI_+(x_0, y)$ is equal to zero, and thus, the hazy input-output map is locally injective around x_0. □

We consider now the sharp input-output map.

THEOREM 6.3. *Let us assume that the graphs of the set-valued maps $F(t, \cdot): X \rightsquigarrow X$ are closed and convex. Let H be a linear operator from X to another finite-dimensional vector-space Y. Let $\bar{x}(\cdot)$ be a solution to differential inclusion* (23). *If the contingent variational inclusion* (37) *is globally sharply observable through H at zero, then the system* (23) *is globally sharply observable through H at $(x_0, H(\bar{x}))$.*

Proof. We apply Proposition 4.1 to the sharp input-output map $I_- := H \circ \mathcal{S}$. We have to prove that the kernel of the algebraic derivative $D_a I_-(x_0, y_0)$ of I_- (where $y_0 := H(\bar{x})$) is equal to zero. Consider $\mathcal{S}$ as a map from $\mathbf{R}^n$ to the Sobolev space $W^{1,1}(0, T; \mathbf{R}^n)$.

Since the graph of the solution map $\mathcal{S}$ is convex (for the graphs of the set-valued maps $F(t, \cdot)$ are assumed to be convex), and since the map H is linear, we know that the chain rule (22) holds true:

$$D_a I_-(x_0, y_0)(u) = (H \circ D_a \mathcal{S}(x_0, \bar{x}(\cdot)))(u). \tag{39}$$

It remains to check that the algebraic derivative $D_a\mathcal{S}(x_0, \bar{x})(u)$ of $\mathcal{S}$ is contained in the subset $\Psi_a(u)$ of solutions to the algebraic variational inclusion starting at u:

$$D_a\mathcal{S}(x_0, \bar{x}(\cdot))(u) \subset \Psi_a(u) := \{w(\cdot) \mid w'(t) \in D_a F(\bar{x}(t), \bar{x}'(t))(w(t)) \text{ and } w(0) = u\}.$$

Since the algebraic derivative of a convex set-valued map is contained in the contingent derivative, then the set $\Psi_a(u)$ is contained in the subset $\Psi(u)$ of solutions to the contingent variational inclusion (37) starting at u. Hence the kernel of $D_a I_-(x_0, y_0)$ is contained in the kernel of $H \circ \Psi$. But to say that the kernel of $H \circ \Psi$ is equal to zero amounts to saying that the contingent variational inclusion (37) is sharply globally observable through H at zero. Therefore the kernel of $D_a I_-(x_0, y_0)$ is equal to zero, and thus, the inverse image of sharp input-output map at y_0 contains a unique element. This concludes the proof. □

If we assume that the chain rule holds true, we can state the following proposition, a consequence of the general principle (Theorem 4.1) and of Theorem 5.3 on the estimate of the contingent derivative of the solution map.

PROPOSITION 6.1. *Let us assume that the solution map of the differential inclusion* (23) *and the differentiable observation map H do satisfy the chain rule*

$$DI_-(x_0, y_0)(u) = (H'(\bar{x}) \circ D\mathcal{S}(x_0, \bar{x}(\cdot)))(u)$$

(*where* $y_0 = H(\bar{x})$). *If the contingent variational inclusion*

$$w'(t) \in \overline{\text{co}}\, DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. } in\ [0, T]$$

is globally sharply observable through $H'(\bar{x}(\cdot))$ *at zero, then the system* (23) *is locally sharply observable through* H *at* $(x_0, H(\bar{x}))$.

However, we can bypass the chain rule formula and attempt to obtain directly other criteria of local sharp observability in the nonconvex case.

THEOREM 6.4. *Assume that* F *has closed convex images, is continuous, Lipschitz in the second variable with a constant independent of* t *and that the growth of* F *is linear with respect to the state. Let* H *be a twice continuously differentiable function from* X *to another finite-dimensional vector-space* Y. *Consider an observation* $y^* \in I_-(x_0)$ *and assume that for every solution* $\bar{x}(\cdot)$ *to the differential inclusion* (23) *satisfying* $y^*(\cdot) = H(\bar{x}(\cdot))$ *and for all* $t \in [0, T]$ *we have*

$$\text{Ker } H'(\bar{x}(t)) \subset (F(t, \bar{x}(t)) - F(t, \bar{x}(t)))^{\perp}. \tag{40}$$

If for all $\bar{x}$ *as above the contingent variational inclusion*

$$w'(t) \in \overline{\text{co}}\, DF(t, \bar{x}(t), \bar{x}'(t))(w(t)) \quad \text{a.e. in } [0, T] \tag{41}$$

is globally sharply observable through $H'(\bar{x}(t))$ *at zero, then the system* (23) *is locally sharply observable through* H *at* (x_0, y^*).

Proof. Assume for a moment that the inclusion (23) is not locally sharply observable through H at (x_0, y^*). Then there exists a sequence $x_0^n \neq x_0$, $x_0^n \to x_0$ such that $y^* \in I_-(x_0^n)$, i.e., for some $x_n \in \mathcal{S}(x_0^n)$

$$y^* = H(x_n(\cdot)). \tag{42}$$

Taking a subsequence if needed and keeping the same notation, we may assume that $x_n \to \bar{x}$ weakly in $W^{1,\infty}(0, T; \mathbf{R}^n)$. Then (42) yields that for every $t \in [0, T]$ where $\bar{x}'(t)$, $x_n'(t)$ do exist

$$H'(\bar{x}(t))\bar{x}'(t) = H'(x_n(t))x_n'(t). \tag{43}$$

We shall prove that the convergence is actually strong and even more, that there exists a constant $c > 0$ such that

$$\|x_n'(t) - \bar{x}'(t)\| \leqq c\|x_n(t) - \bar{x}(t)\| \quad \text{a.e. in } [0, T]. \tag{44}$$

Indeed otherwise there exist sequences t_k and n_k such that

$$x_{n_k}'(t_k) \in F(t_k, x_{n_k}(t_k)), \qquad \bar{x}'(t_k) \in F(t_k, \bar{x}(t_k)),$$

$$\|x_{n_k}'(t_k) - \bar{x}'(t_k)\| \geqq k\|x_{n_k}(t_k) - \bar{x}(t_k)\|.$$

Taking a subsequence and keeping the same notation, by continuity of F, we may assume that for some $t \in [0, T]$, $p \in F(t, \bar{x}(t))$

$$t_k \to t, \qquad \bar{x}'(t_k) \to p. \tag{45}$$

Let ρ denote the Lipschitz constant of F with respect to x and let $y(t_k) \in F(t_k, \bar{x}(t_k))$ be such that

$$\|y(t_k) - \bar{x}_{n_k}'(t_k)\| \leqq \rho\|x_{n_k}(t_k) - \bar{x}(t_k)\|. \tag{46}$$

Since H' is locally Lipschitz and x'_{n_k} are equibounded, from the last inequality and (43) we deduce that for some constants M, $M_1 > 0$

$$
\begin{aligned}
&\|H'(\bar{x}(t_k))(y(t_k) - \bar{x}'(t_k))\| \\
&\quad \leqq \|H'(\bar{x}(t_k))(x'_{n_k}(t_k) - \bar{x}'(t_k))\| + \rho\|H'(\bar{x}(t_k))\|\,\|x_{n_k}(t_k) - \bar{x}(t_k)\| \\
&\quad \leqq \|H'(x_{n_k}(t_k))x'_{n_k}(t_k) - H'(\bar{x}(t_k))\bar{x}'(t_k)\| + M\|x_{n_k}(t_k) - \bar{x}(t_k)\|\,\|x'_{n_k}(t_k)\| \\
&\qquad + \rho\|H'(\bar{x}(t_k))\|\,\|x_{n_k}(t_k) - \bar{x}(t_k)\| \leqq M_1\|x_{n_k}(t_k) - \bar{x}(t_k)\|.
\end{aligned} \tag{47}
$$

From (46) and the choice of t_k, we obtain

$$
\frac{\|y(t_k) - \bar{x}'(t_k)\|}{\|x_{n_k}(t_k) - \bar{x}(t_k)\|} \to \infty \quad \text{when } k \to \infty. \tag{48}
$$

It is also not restrictive to assume that for some u of $\|u\| = 1$

$$
u_k := \frac{y(t_k) - \bar{x}'(t_k)}{\|y(t_k) - \bar{x}'(t_k)\|} \to u. \tag{49}
$$

Then (47), (48) yield

$$
u \in \operatorname{Ker} H'(\bar{x}(t)).
$$

On the other hand, u_k is contained in the space spanned by $F(t_k, \bar{x}(t_k)) - F(t_k, \bar{x}(t_k))$ and, by continuity of F, u is contained in the space spanned by $F(t, \bar{x}(t)) - F(t, \bar{x}(t))$. Since $u \neq 0$ this contradicts (40) and therefore (44) follows.

From the Gronwall inequality and (44) we deduce that for some $M_2 > 0$

$$
\|x_n(t) - \bar{x}(t)\| \leqq M_2\|x_n(0) - \bar{x}(0)\|.
$$

Setting $h_n = \|x_0^n - x_0\|$, we obtain

$$
\|x_n - \bar{x}\|_{W^{1,\infty}(0,T)} \leqq M_2 h_n.
$$

Taking a subsequence and keeping the same notation, we may assume that

$$
\frac{x_n - \bar{x}}{h_n} \to w \quad \text{weakly in } W^{1,\infty}(0, T).
$$

By Theorem 5.3, w is a solution of the contingent variational inclusion of (41). Moreover, $w(0) \neq 0$. Since $H(x_n(\cdot)) = H(\bar{x}(\cdot))$ taking the derivatives we obtain that for every $t \in [0, T]$, $H'(\bar{x}(t))w(t) = 0$. This contradicts the assumptions of Theorem 6.4 and completes the proof. □

Example. Observability around an equilibrium. Let us consider the case of a time-independent system (F, H): this means that the set-valued map $F: X \rightsquigarrow X$ and the observation map $H: X \rightsquigarrow Y$ do not depend on the time.

We shall observe this system around an equilibrium $\bar{x}$ of F, i.e., a solution to the equation

$$
0 \in F(\bar{x}). \tag{50}
$$

For simplicity, we assume that the set-valued map F is **sleek** at the equilibrium. Hence all the derivatives of F at $(\bar{x}, 0)$ do coincide with the contingent derivative $DF(\bar{x}, 0)$, which is a closed convex process from X to itself.

The theorems on local observability reduce the local observability around the equilibrium $\bar{x}$ to the study of the observability properties of the variational inclusion

$$
w'(t) \in DF(\bar{x}, 0)(w(t)) \tag{51}
$$

through the observation map $H'(\bar{x})$ around the solution zero of this variational inclusion.

We mention below a characterization of sharp observability of the variational inclusion in terms of "viability domains" of the restriction of the derivative $DF(\bar{x}, 0)$ to the kernel of $H'(\bar{x})$.

Recall that a subset $P \subset \ker H'(\bar{x})$ is a "viability domain" if

$$\forall w \in P, \quad DF(\bar{x}, 0)(w) \cap T_P(w) \neq \varnothing$$

where $T_P(w)$ denotes the "contingent cone to P at $w \in P$."

PROPOSITION 6.2. *Let us assume that F is sleek at its equilibrium $\bar{x}$ and that H is differentiable at $\bar{x}$. Then the variational inclusion* (51) *is sharply observable through $H'(\bar{x})$ at zero if and only if the* **largest closed viability domain** *of the restriction to* $\ker H'(\bar{x})$ *of the contingent derivative $DF(\bar{x}, 0)$ is equal to zero.*

Proof. Let us denote by E the restriction of the contingent derivative $DF(\bar{x}, 0)$ to the kernel of $H'(\bar{x})$ defined by

$$E(u) := \begin{cases} DF(\bar{x}, 0)(u) & \text{if } u \in \ker H'(\bar{x}), \\ \varnothing & \text{if } u \notin \ker H'(\bar{x}). \end{cases} \tag{52}$$

We consider the associated differential inclusion

$$w'(t) \in E(w(t)). \tag{53}$$

We know that the **largest closed viability domain** of the closed convex process E is the domain of the solution map of the associated differential inclusion (53). (See [6] and [7].)

But if we denote by $\mathcal{R}$ the solution map of the variational inclusion (51) and by $\mathcal{B}$ the set of functions $x(\cdot)$ such that

$$\forall t \in [0, T], \quad x(t) \in \ker H'(\bar{x}),$$

then we observe that the solution map of the differential inclusion (53) is the set-valued map $u \rightsquigarrow \mathcal{R}(u) \cap \mathcal{B}$. Hence its domain is the set $\mathcal{R}^{-}(\mathcal{B})$. Since

$$\mathcal{R}^{-}(\mathcal{B}) = \ker (H'(\bar{x}) \circ \mathcal{R}),$$

we infer that **the largest viability domain of E is the kernel of the sharp input–output map** $H'(\bar{x}) \circ \mathcal{R}$.

Consequently, the variational inclusion (51) being sharply observable if and only if the kernel of $H'(\bar{x}) \circ \mathcal{R}$ is equal to zero, our proposition ensues. □

Remark. In the same way, the variational inclusion (51) is hazily observable if and only if the kernel of $H'(\bar{x}) \square \mathcal{R}$ is equal to zero.

There are also some relations between the kernel of the hazy input-output map $H'(\bar{x}) \square \mathcal{R}$ and the largest invariance domain of the restriction of the derivative to the kernel of $H'(\bar{x})$. First, we remark that

$$\mathcal{R}^{+}(\mathcal{B}) = \ker (H'(\bar{x}) \square \mathcal{R}),$$

i.e., that the kernel of $H'(\bar{x}) \square \mathcal{R}$ is the largest set enjoying the "invariance property": for any $u \in \ker H'(\bar{x}) \square \mathcal{R}$, all solutions to differential inclusion (53) remain in this kernel.

When F is Lipschitzian in a neighborhood of $\ker H'(\bar{x})$, any closed subset $P \subset \ker H'(\bar{x})$ that is "invariant" in the sense that

$$\forall w \in P, \quad DF(\bar{x}, 0)(w) \subset T_P(w)$$

enjoys the invariance property. The converse is true only if we assume that the domain of $DF(\bar{x}, 0)$ is the whole space.

Then, if such is the case, *the variational inclusion is hazily observable through* $H'(\bar{x})$ *at zero if and only if the largest closed invariance domain of the restriction to* $\ker H'(\bar{x})$ *of the derivative* $DF(\bar{x}, 0)$ is equal to zero.

Remark. We have proved in [5] that under some further conditions the sharp observability of the variational inclusion at zero is equivalent to the controllability of the adjoint system

$$-p'(t) \in DF(\bar{x}, 0)^*(p(t)) + H'(\bar{x})^* u(t), \qquad u(t) \in Y^*. \tag{54}$$

PROPOSITION 6.3. *We posit the assumptions of Proposition* 6.2, *we assume that* $DF(\bar{x}, 0)(0) = 0$ *and we suppose that*

$$\ker H'(\bar{x}) + \mathrm{Dom}\,(DF(\bar{x}, 0)) = X. \tag{55}$$

Then the sharp observability at zero of the variational inclusion (51) *is equivalent to the controllability of the adjoint system* (54).

(About eleven characterizations of this property are supplied in [5].)

Proof. Assumption (55) implies that the transpose E^* of the restriction E of the closed convex process $DF(\bar{x}, 0)$ to $\ker H'(\bar{x})$ is given by the formula

$$(DF(\bar{x}, 0)|_{\ker H'(\bar{x})})^* = DF(\bar{x}, 0)^* + \mathrm{Im}\,(H'(\bar{x})^*) \tag{56}$$

(see [2, Cor. 3.3.17, p. 142]).

We also know (see [5, Prop. 1.12, p. 1198]) that if the domain of the transpose E^* of E is the whole space, then a vector subspace P is an invariance domain of E if and only if its orthogonal $P^\perp$ is a viability domain of E^* (this is also true when the domain of E is the whole space. But this does not apply to our case, since the domain of E is the kernel of $H'(\bar{x})$). Since the domain of E^* is equal to the domain of $DF(\bar{x}, 0)^*$ (thanks to formula (56)), this condition is equivalent to $DF(\bar{x}, 0)(0) = 0$.

Hence the variational inclusion (51) being sharply observable at zero if and only if the largest closed viability domain of E is equal to zero (by Proposition 6.2), we deduce that this happens if and only if the smallest invariance domain of E^* is equal to X, i.e., if and only if the adjoint system (54) is controllable.

Therefore, our statement follows from Theorem 5.5 of [5, p. 1207]. □

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Grundlehren der Mathematischen Wissenschaften 264, Springer-Verlag, Berlin, New York, 1984.

[2] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.

[3] J.-P. AUBIN AND H. FRANKOWSKA, *On inverse function theorems for set-valued maps*, J. Math. Pure Appl., 66 (1987), pp. 71-89.

[4] ———, *Set-valued analysis*, monograph, to appear.

[5] J.-P. AUBIN, H. FRANKOWSKA, AND C. OLECH, *Controllability of convex processes*, SIAM J. Control Optim., 24 (1986), pp. 1192-1211.

[6] J.-P. AUBIN, *Smooth and heavy solutions to control problems*, in Proc. Conference on Functional Analysis, Santa Barbara, CA, June, 1985, Marcel Dekker, New York, Basel, pp. 24-26.

[7] ———, *Viability theory*, monograph, to appear.

[8] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247-262.

[9] S. DOLECKI AND D. L. RUSSEL, *A general theory of observation and control*, SIAM J. Control Optim., 15 (1977), pp. 185-220.

[10] H. FRANKOWSKA, *Local controllability of control systems with feedback*, J. Optim. Theory Appl., 60 (1989), pp. 277-296.

[11] ———, *Local controllability and infinitesimal generators of semi-groups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412-432.

[12] ———, *The maximum principle for an optimal solution to a differential inclusion with end point constraints*, SIAM J. Control Optim., 25 (1987), pp. 145-157.

[13] ———, *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim., to appear.

[14] G. HADDAD, *Monotone trajectories of differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 83-100.
[15] R. HERMANN AND A. J. KENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, 22 (1977), pp. 728-740.
[16] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Lecture Notes in Control and Information Sciences, 72, Springer-Verlag, Berlin, New York, 1985.
[17] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47-52.
[18] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., (1985), pp. 197-216.
[19] A. B. KURZHANSKII, *Control and Observation under Conditions of Uncertainty*, Nauka, Russia, 1977.
[20] G. LEITMANN, *The Calculus of Variations and Optimal Control*, Plenum Press, New York, 1981.
[21] R. T. ROCKAFELLAR, *Monotone Processes of Convex and Concave Type*, Mem. Amer. Math. Soc., Providence, RI, 1967.
[22] ———, *La Théorie des Sous-Gradients*, Presses de l'Université de Montréal, Montréal, Quebec, Canada, 1979.
[23] SHI SHUZHONG, *Choquet Theorem and Nonsmooth Analysis*, Cahiers Math. Décision, (1987), #8621.
[24] ———, *Théorème de Choquet et analyse non régulière*, C.R. Acad. des Sci. Paris, 305 Sér. 1, pp. 41-44.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 976-990, September 1989

005

GLOBAL DIRECTIONAL CONTROLLABILITY*

J. WARGA†

This paper is dedicated to the memory of E. J. McShane.

Abstract. This paper derives sufficient conditions for the image of a function $\varphi_1: Q \to \mathbb{R}^m$ to cover an open segment originating at $\varphi_1(\bar{q})$ and with direction w (or even a neighborhood of such an interval) subject to restrictions of the form $\varphi_2(q) \in C$ or $\varphi_2(q) + G \subset C$. Here $\varphi_2: Q \to \mathscr{Z}$, Q is an arbitrary set, C a convex subset of the topological vector space $\mathscr{Z}$, and G a neighborhood of 0 in $\mathscr{Z}$. Also derived are similar directional controllability conditions subject to the additional restriction $q \in \mathscr{U} \subset Q$ for an "abundant" subset $\mathscr{U}$ of Q. These conditions are applicable to unilateral problems of control theory and of (infinite-dimensional) mathematical programming. These results are global and apply to problems defined by functions whose restrictions to certain finite-dimensional sets are differentiable (but not necessarily C^1) or are locally uniform limits of differentiable functions.

Key words. directional controllability, conical controllability, sufficient conditions, inclusion restrictions, attainable sets, nonsmooth data

AMS(MOS) subject classifications. 45E15, 49E30

1. Introduction.

The concepts of controllability originally arose in the study of linear control systems and their attainable sets [1], [12] and were later applied to smooth nonlinear systems [3], [4]. In our own studies (e.g., in [6]-[9]) we had adapted this concept to nonlinear and nonsmooth systems subject to infinite-dimensional inclusion restrictions. A model of such control systems is provided by the objects

$$\bar{q} \in Q, \quad \mathscr{U} \subset Q, \quad C \subset \mathscr{Z}, \quad \varphi = (\varphi_1, \varphi_2): Q \to \mathbb{R}^m \times \mathscr{Z},$$

where Q is an arbitrary set (but usually assumed to be a convex subset of a real vector space), $\mathscr{U}$ an "abundant" subset of Q, C a convex subset of a topological vector space $\mathscr{Z}$, and $\varphi_2(\bar{q}) \in C$. In the context of control theory, $\mathscr{U}$ may represent the set of ordinary control functions (or a subset closed under measurable concatenations), Q the corresponding set of relaxed controls, φ_1 an m-dimensional function determined by the solution $x(\cdot)$ of the (differential or functional-integral) equation of motion controlled by $u \in \mathscr{U}$ or $q \in Q$, and the restriction $\varphi_2(q) \in C$ a unilateral state constraint such as $x(t) \in A$ for all t.

Crudely speaking, we say that φ_1 is controllable at $\bar{q}$ if $\varphi_1(\mathscr{A})$ covers a neighborhood of $\varphi_1(\bar{q})$, where $\mathscr{A}$ is the set $\mathscr{U} \cap \varphi_2^{-1}(C)$ or, in a stronger type of controllability,

$$\mathscr{A} = \{u \in \mathscr{U} \mid \varphi_2(u) + G \subset C\}$$

for some neighborhood G of 0 in $\mathscr{Z}$. Typically, controllability is studied at points $\bar{q}$ that are extremals (or stationary points) of the problem because controllability rules out $\bar{q}$ as a solution of a related optimization problem. However, even if φ_1 is not (or cannot be shown to be) controllable at $\bar{q}$, it is still of interest to determine whether $\varphi_1(\mathscr{A})$ extends in the direction w from $\varphi_1(\bar{q})$, i.e., whether $\varphi_1(\mathscr{A})$ covers an open segment originating at $\varphi_1(\bar{q})$ and with direction w or even a neighborhood of such a segment. We had initiated the study of such directional controllability in a previous paper [11] which deals with higher-order conditions for local conical controllability, the term "conical" referring to the shape of the neighborhood. In the present paper we study first-order conditions for global directional controllability without smoothness

* Received by the editors March 7, 1988; accepted for publication November 1, 1988. This research was partially supported by National Science Foundation grant DMS-8619002.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

assumptions ("smoothness" being the term used in nonsmooth analysis to mean "continuous differentiability" and not necessarily C^∞ behavior).

We present the basic results in § 2, some examples in § 3, and the proofs in § 4.

2. Covering theorems. We shall assume given the objects $\bar{q}$, Q, $\mathcal{U}$, C, $\mathcal{Z}$, φ as described in § 1. We endow each space $\mathbb{R}^s$ for $s = 1, 2, \cdots$ with the norm $|(x_1, \cdots, x_s)| = \sum_{i=1}^{s} |x_i|$, and denote by B_s and $\bar{B}_s$ the open and closed unit balls in $\mathbb{R}^s$, by $d[p, A]$ the distance of the point p to the set A, and by A°, $\bar{A}$, and co A the interior, the closure, and the convex hull of a set A. We refer to a set B as a neighborhood of a point x, respectively, of a set A if $x \in B^\circ$, respectively, $A \subset B^\circ$. For A, $B \subset \mathcal{Z}$, we write

$$A \ominus B := \{z \in \mathcal{Z} \mid z + B \subset A\}.$$

If W is a convex subset of $\dot{\mathbb{R}}^k$ with $W^\circ \neq \varnothing$, $p: W \to \mathbb{R}^m \times \mathcal{Z}$, and $x \in W$ (but x is not necessarily in W°) we say that p is differentiable at x and has the derivative $p'(x)$ [5, p. 167] if $p'(x)$ is a linear operator from $\mathbb{R}^k$ to $\mathbb{R}^m \times \mathcal{Z}$ such that

$$\lim |\xi - x|^{-1}[p(\xi) - p(x) - p'(x)(\xi - x)] = 0 \quad \text{as } \xi \to x, \quad \xi \in W \backslash \{x\}.$$

Condition 2.0. Let $X \subset \mathbb{R}^k$ and $\hat{q}: X \to Q$. We say that φ and $\hat{q}$ satisfy Condition 2.0 if for every $x \in X$ there exists a sequence $(u_n(x))$ in $\mathcal{U}$ such that

$$\lim_n \varphi(u_n(x)) = \varphi(\hat{q}(x)) \quad \text{uniformly for } x \in X$$

and

$$x \to \varphi(u_n(x)): X \to \mathbb{R}^m \times \mathcal{Z} \text{ is continuous for each } n = 1, 2, \cdots.$$

Remark. This definition is motivated by problems of optimal control in which $\mathcal{U}$ represents an "abundant" set of ordinary controls (e.g., a set closed under measurable concatenations) and Q the corresponding set of relaxed controls. If φ is continuous (with respect to the weak star topology of relaxed controls [5, Chap. IV]),

$$X = \left\{ \theta = (\theta_1, \cdots, \theta_k) \mid \theta_j \geqq 0, \sum_{i=1}^{k} \theta_i \leqq 1 \right\}, \quad \bar{q} \in Q, \quad y_j \in Q - \bar{q},$$

and

$$\hat{q}(\theta) = \bar{q} + \sum_{j=1}^{k} \theta_j y_j,$$

then φ and $\hat{q}$ satisfy Condition 2.0 [5, Thm. IV.3.9, p. 285].

THEOREM 2.1. *Let $w \in \mathbb{R}^m$, $\alpha_0 > 0$, X be a closed subset of $\mathbb{R}^k$, $0 \in X$, G an open neighborhood of 0 in $\mathcal{Z}$, $\mathcal{T} \subset \mathbb{R}^k$ compact and convex, and $\psi = (\psi_1, \psi_2): X \to \mathbb{R}^m \times \mathcal{Z}$.*

Assume that

(a) $\psi_2(0) \in C$,

(b) *ψ is differentiable at every $x \in X$,*

(c) *for each $x \in X \cap [0, \alpha_0]\mathcal{T}$ there exist $\eta(x) \in \mathcal{T}$ and $r(x)$, $r_1(x) > 0$ such that*

$$\psi'(x)\eta(x) \in \{w\} \times [C \ominus G - \psi_2(x)]$$

$$\eta(x) + r(x)\bar{B}_k \subset \mathcal{T}, \qquad x + [0, r_1(x)][\eta(x) + r(x)\bar{B}_k] \subset X.$$

Then

$$(*) \quad \psi_1(0) + \alpha w \in \{\psi_1(x) \mid x \in X \cap \alpha\mathcal{T}, \psi_2(x) + (1 - e^{-\alpha})G \subset C\} \qquad \forall \alpha \in [0, \alpha_0].$$

If, furthermore,

(d) $\varphi: Q \to \mathbb{R}^m \times \mathscr{Z}$ *and* $\hat{q}: X \to Q$ *satisfy Condition* 2.0, $\bar{q} = \hat{q}(0)$, *and* $\psi(x) = \varphi(\hat{q}(x))$ *for all* $x \in X$,
and

(e) $\psi_1'(x)$ *is of full rank for all* $x \in X \cap [0, \alpha_0]\mathscr{T}$
then for every $\alpha \in (0, \alpha_0)$ *and* $s \in (0, 1)$, *there exists a neighborhood* $V_{s,\alpha}$ *of* 0 *in* $\mathbb{R}^m$ *such that*

$$(**) \qquad \varphi_1(\bar{q}) + \alpha w + V_{s,\alpha} \subset \{\varphi_1(u) \,|\, u \in \mathscr{U},\ \varphi_2(u) + (1 - e^{-\alpha}) sG \subset C\}.$$

If condition (d) *holds and*
(f) *the functions*

$$\rho: [0, \alpha_0] \to (0, 1], \quad \beta: (0, \alpha_0] \to [0, \infty), \quad \delta: [0, \alpha_0] \to (0, 1)$$

are pointwise limits from below of positive continuous functions and such that, for all $\alpha \in [0, \alpha_0]$ *and* $x \in X \cap \alpha\mathscr{T}$,

$$\rho(\alpha) \leqq r(x), \quad \psi_1'(x)\bar{B}_k \supset \beta(\alpha)\bar{\beta}_m, \quad r(\alpha)\psi_2'(x)\bar{B}_k \subset [1 - \delta(\alpha)]G$$

then, setting

$$I(\alpha) = \int_0^\alpha \rho(t)\beta(t)\, dt, \qquad J(\alpha) = \int_0^\alpha e^{-(\alpha - t)}\delta(t)\, dt,$$

we have

$$(***) \qquad \varphi_1(\bar{q}) + \alpha w + I(\alpha)\bar{B}_m \subset \{\varphi_1(q) \,|\, q \in Q,\ \varphi_2(q) + J(\alpha)G \subset C\}$$

$$\forall \alpha \in [0, \alpha_0]$$

and

$$(****) \qquad \varphi_1(\bar{q}) + \alpha w + I(\alpha)\bar{B}_m \subset \{\varphi_1(u) \,|\, u \in \mathscr{U},\ \varphi_2(u) + sJ(\alpha)G \subset C\}$$

$$\forall \alpha \in (0, \alpha_0),\ s \in (0, 1).$$

Remarks.

1. A special case of Theorem 2.1 (and of Theorem 2.2 below) applies to problems without inclusion restrictions. If $\varphi = \varphi_1: Q \to \mathbb{R}^m$ and $\psi = \psi_1: X \to \mathbb{R}^m$ are given, then we can apply Theorem 2.1 by setting

$$\mathscr{Z} = C = \mathbb{R}, \quad \varphi_2(q) = \psi_2(x) = 0 \quad \forall q \in Q, x \in X.$$

Then all the assumptions referring to C, $\mathscr{Z}$, φ_2, ψ_2 are automatically satisfied.

2. We observe that the last two conditions in (c) are automatically satisfied for appropriate choices of $r(x)$, $r_1(x)$ if X is a cone, $\mathscr{T} \subset X$ and $\eta(x) \in \mathscr{T}^\circ$ for all $x \in \alpha_0\mathscr{T}$.

3. In Theorem 2.2 below we drop the assumption of Theorem 2.1 that the function ψ is differentiable. We assume, instead, that ψ can be locally uniformly approximated by differentiable functions that essentially satisfy the conditions of Theorem 2.1. This approach is related to the previously introduced concepts of derivate containers [6]-[8], [10] and of Frankowska's $\mathscr{P}_b$-set [2] but, in the present case, it requires somewhat weaker hypotheses.

THEOREM 2.2. *Let* $w \in \mathbb{R}^m$, G *be a neighborhood of* 0 *in* $\mathscr{Z}$, X *a closed subset of* $\mathbb{R}^k$, $0 \in X$, $\mathscr{T} \subset \mathbb{R}^k$ *compact and convex,* $\alpha_0 > 0$, *and* $\hat{q}: X \to Q$ *such that the function*

$$x \to \psi(x) := \varphi(\hat{q}(x)): X \to \mathbb{R}^m \times \mathscr{Z}$$

is continuous, and $\psi_2(0)=\varphi_2(\bar{q})\in C$, *where* $\bar{q}:=\hat{q}(0)$. *Assume that for each* $\bar{x}\in X\cap[0,\alpha_0]\mathcal{T}$ *there exist a closed neighborhood* $W(\bar{x})$ *of* $\bar{x}$ *and differentiable functions* $\psi^i=(\psi_1^i,\psi_2^i):X\cap W(\bar{x})\to\mathbb{R}^m\times\mathcal{Z}$ *for all* $i=1,2,\cdots$ *such that*

$$\lim_i \psi^i=\psi \text{ uniformly on } X\cap W(\bar{x})$$

and each ψ^i *satisfies condition* (c) *of Theorem* 2.1. *Then*

$$\varphi_1(\bar{q})+\alpha w\in\{\psi_1(x)\,|\,x\in X\cap\alpha\mathcal{T},\ \psi_2(x)+(1-e^{-\alpha})G\subset C\}$$
$$\subset\{\varphi_1(q)\,|\,q\in Q,\ \varphi_2(q)+(1-e^{-\alpha})G\subset C\}\quad\forall\alpha\in[0,a_0].$$

3. Examples. In the two examples below there are no inclusion restrictions.

Example I. Let

$$Q=\mathcal{U}=\mathbb{R}^2,\quad \varphi=(\varphi^1,\varphi^2),\quad x=(x_1,x_2),$$
$$\varphi^1(x)=x_1+|x_1+x_2|^{1/2},\qquad \varphi^2(x)=x_1-x_2.$$

We choose an arbitrary $\alpha_0\geqq 1$ and set

$$X=\{x\in\mathbb{R}^2\,|\,x_1+x_2\geqq 0\},\qquad \mathcal{T}=X\cap\bar{B}_2$$
$$\hat{q}(x)=x,\quad \psi=\varphi|_X,\quad \eta(x)=(\eta_1,\eta_2)(x).$$

In the neighborhood of any $x\in X$, we approximate ψ by functions $\chi_j(=\chi_j^1,\chi_j^2)$, defined by

$$\chi_j^1(y)=y_1+h_j(y_1+y_2),\qquad \chi_j^2(y)=y_1-y_2,$$

where $h_j(z)$ can be any increasing differentiable functions that uniformly approximate $z^{1/2}$ on $[0,1]$ as $j\to\infty$, e.g.,

$$h_j(z)=\tfrac{3}{4}j^{-1/2}+\tfrac{1}{4}j^{3/2}z^2\quad\text{for } 0\leqq z\leqq 1/j,\qquad h_j(z)=z^{1/2}\quad\text{for } z>1/j.$$

For $w\in\mathbb{R}^2$, the equation $x_j'(x)\eta(x)=w$ is of the form

$$(1+a)\eta_1(x)+a\eta_2(x)=w_1,\qquad \eta_1(x)-\eta_2(x)=w_2,$$

where $a=h_j'(x_1+x_2)$. This equation has the solution

(1) $$\eta_1(x)=(1+2a)^{-1}[w_1+aw_2],\qquad \eta_2(x)=(1+2a)^{-1}[w_1-(1+a)w_2].$$

Since $a\geqq 0$, it follows that

(2) $$|\eta(x)|=|\eta_1(x)|+|\eta_2(x)|\leqq|w_2|+2|w_1|.$$

Thus, by (1) and (2), $\eta(x)\in\mathcal{T}^\circ$ (i.e., $\eta_1(x)+\eta_2(x)>0$ and $|\eta(x)|<1$) if

$$w\in W:=\{v=(v_1,v_2)\in\mathbb{R}^2\,|\,2v_1-v_2>0,\ |v_2|+2|v_1|<1\}.$$

Since X is a convex cone and $\mathcal{T}\subset X$, it follows, by Theorem 2.2, that for all $a_0>1$ and $w\in W$, we have

(3) $$\psi(0)+\alpha w=\alpha w\in\psi(\alpha\mathcal{T})\quad\forall\alpha\in[0,\alpha_0].$$

Since $\alpha\mathcal{T}$ is compact, relation (3) remains valid for all $w\in\bar{W}$ and, since α_0 can be taken arbitrarily large, we have

(4) $$\psi(\{x\,|\,x_1+x_2\geqq 0\})\supset\{v=\{v_1,v_2)\,|\,2v_1-v_2\geqq 0\}.$$

We can verify that this result is the best possible by setting

$$\zeta_1=x_1+x_2,\qquad \zeta_2=x_1-x_2$$

which transforms the equation $\psi(x)=v=(v_1,v_2)$ into

(5) $$\zeta_1+2|\zeta_1|^{1/2}=2v_1-v_2,\qquad \zeta_2=v_2.$$

On the other hand, (5) can also be solved when $2v_1 - v_2 < 0$ but that requires that we restrict ourselves to values of $\zeta_1 = x_1 + x_2 < -4$. This suggests that we might proceed as above but with

$$X = \{x \in \mathbb{R}^2 \,|\, x_1 + x_2 \leqq 0\}.$$

However, as will be seen below, Theorem 2.2 is based on Theorem 2.1, which is derived by a procedure bearing some similarity to an approximate integration of the differential equation $dx/d\alpha = \eta(x)$ with the initial condition $x(0) = 0$ (a differential equation which is unconventional because its right-hand side may be discontinuous). Therefore, this theorem cannot yield results requiring an immediate jump from 0 to $x_1 + x_2 < -4$. However, if we set

$$y_1 = -2 - x_1, \qquad y_2 = -2 - x_2$$

and use (y_1, y_2) as the new variable, with X and $\mathcal{T}$ defined as before, then we can use Theorem 2.1 to show that

$$\varphi(\mathbb{R}^2) \supset \{v = (v_1, v_2) \,|\, 2v_1 - v_2 \leqq 0\}$$

and therefore, by combination with relation (4), that $\varphi(\mathbb{R}^2) = \mathbb{R}^2$.

Example II. Let $Q = \mathcal{U} = \mathbb{R}^3$, $x = (x_1, x_2, x_3)$,

$$\varphi^1(x) = |x_2 - x_1^2| + x_3, \qquad \varphi^2(x) = x_1 + x_2 + x_3.$$

We at first choose some $R \geqq 1$ and set

$$X = X_1 = \{x \in \mathbb{R}^3 \,|\, x_2 \leqq 0\},$$

$$\mathcal{T} = \mathcal{T}_1 = \{x \in \mathbb{R}^3 \,|\, x_2 \leqq 0, |x_1| + |x_2|/R + |x_3|/R \leqq 1\} \subset X_1,$$

$$\alpha_0 = \tfrac{1}{4}, \quad \hat{q}(x) = x, \quad \psi = \varphi|_{X_1}, \quad \eta(x) = (\eta_1, \eta_2)(x).$$

For each $x \in X$ and $w = (w_1, w_2) \in \mathbb{R}^2$, the equation $\psi'(x)\eta = w$ has the form

$$\eta_1(2x_1, 1) + \eta_2(-1, 1) + \eta_3(1, 1) = (w_1, w_2).$$

This equation will have a solution $\eta(x)$ in $\mathcal{T}^\circ$ for all $x \in X \cap [0, \frac{1}{4}]\mathcal{T}$—and $\eta(x)$ will thus satisfy the relations

$$\eta(x) + r(x)\bar{B}_k \subset \mathcal{T}, \qquad x + [0, r_1(x)][\eta(x) + r(x)\bar{B}_k] \subset X$$

for appropriate $r(x)$, $r_1(x) > 0$—if $w \in \psi'(x)\mathcal{T}^\circ$ for all $x \in X \cap [0, \frac{1}{4}]\mathcal{T}$. It is easily seen that this is the case if w is in the interior of

$$P_R = \operatorname{co}\{(\tfrac{1}{2}, 1), R(1, -1), R(1, 1), R(-1, -1)\} \subset X.$$

Thus, by relation (*) of Theorem 2.1, for every choice of $R \geqq 1$, $\alpha \in [0, \frac{1}{4}]$, and $w \in P_R$, we have

$$\alpha w = \psi(0) + \alpha w \in \psi(X \cap [0, \tfrac{1}{4}]\mathcal{T}).$$

Therefore,

$$\psi(X_1) \supset P := \tfrac{1}{4} \bigcup_{R \geqq 1} P_R = \{(v_1, v_2) \in \mathbb{R}^2 \,|\, v_2 < v_1 + \tfrac{1}{8}\}. \tag{2}$$

We next consider the complement of P. To do so, we redefine X, $\mathcal{T}$, and ψ as

$$X = X_2 := \{x \in \mathbb{R}^3 \,|\, x_2 \geqq x_1^2, x_1 \geqq 0\}, \qquad \mathcal{T} = \mathcal{T}_2 := 5\bar{B}_3, \qquad \psi = \psi|_{X_2}.$$

For each $x \in X$ and $w = (w_1, w_2) \in \mathbb{R}^2$, the equation $\psi'(x)\eta = w$ now has the form

$$\eta_1(-2x_1, 1) + (\eta_1 + \eta_2)(1, 1) = (w_1, w_2). \tag{3}$$

Assume that $|w| \leqq 1$ and $w_2 - w_1 > 0$, and set

$$\eta_1 = (2x_1+1)^{-1}(w_2 - w_1), \qquad \eta_2 = 3x_1(2x_1+1)^{-1}(w_2 - w_1),$$
$$\eta_3 = (2x_1+1)^{-1}(w_1 + 2x_1 w_2) - \eta_2 \qquad \text{if } x \in X,\ x_1 > 0,$$
$$\eta_1 = w_2 - w_1, \quad \eta_2 = 1, \quad \eta_3 = w_1 - 1 \quad \text{if } x \in X,\ x_1 = 0.$$

Then $\eta(x) = (\eta_1, \eta_2, \eta_3) \in \mathcal{T}$ and $\eta(x)$ is a solution of (3) satisfying the relations $\eta_1 > 0$, $\eta_2 > 2x_1\eta_1$. Thus $\psi(x)$ and $\eta(x)$ satisfy conditions (a)-(c) of Theorem 2.1 for every choice of $\alpha_0 > 0$. It follows that

$$\alpha w = \psi(0) + \alpha w \in \psi(X_2) \quad \text{if } \alpha \geqq 0,\ |w| \leqq 1 \text{ and } w_2 - w_1 > 0$$

so that $\psi(X_2) \supset \{(v_1, v_2) \mid v_2 - v_1 > 0\}$. Together with relation (2) this shows that $\varphi(X_1 \cup X_2) = \mathbb{R}^2$.

4. Proofs.

LEMMA 4.1. *Let A, B, B_1, B_2, $G \subset \mathcal{X}$, and G be an open neighborhood of* 0. Then

(a) $(A \ominus B_1) \ominus B_2 = A \ominus (B_1 + B_2) = (A \ominus B_2) \ominus B_1$

and

(b) $A \ominus G = A^\circ \ominus G = \overline{A \ominus G}$.

If A is convex then

(c) $A \ominus B = A \ominus \operatorname{co} B$ *and this set is convex,*

(d) $\bar{A} \ominus G = A^\circ \ominus G = A \ominus G$,

(e) $A \ominus \lambda_2 G \subset A \ominus \lambda_1 G$ *if* $0 \leqq \lambda_1 \leqq \lambda_2$,

and

(f) $(A + B) \ominus B = A$ *if A is closed and either $A^\circ \neq \varnothing$ or $\mathcal{X}$ is finite-dimensional.*

Proof. We have

$$(A \ominus B_1) \ominus B_2 + B_1 + B_2 = [(A \ominus B_1) \ominus B_2 + B_2] + B_1$$
$$\subset A \ominus B_1 + B_1 \subset A;$$

hence,

$$(A \ominus B_1) \ominus B_2 \subset A \ominus (B_1 + B_2).$$

Conversely, since

$$A \ominus (B_1 + B_2) + B_1 + B_2 \subset A$$

we have

$$A \ominus (B_1 + B_2) + B_2 \subset A \ominus B_1.$$

Therefore,

$$A \ominus (B_1 + B_2) \subset (A \ominus B_1) \ominus B_2,$$

which proves (a).

Let $\bar{x} \in \overline{A \ominus G}$ and $g \in G$. Then $g + V + V \subset G$ for some symmetric neighborhood V of 0. We can find $x \in A \ominus G$ such that $\bar{x} \in x + V$, whence it follows that

$$\bar{x} + g + V \subset x + g + V + V \subset x + G \subset A,$$

implying that $\bar{x} + g \in A^\circ$. Since this is valid for every $g \in G$, we have $\bar{x} + G \subset A^\circ$, hence $\bar{x} \in A^\circ \ominus G$. Thus

$$\overline{A \ominus G} \subset A^\circ \ominus G \subset A \ominus G,$$

which proves (b).

Now assume that A is convex. Then $z \in A \ominus B$ implies that $z + B \subset A$, hence $z + \operatorname{co} B \subset A$ and $z \in A \ominus \operatorname{co} B$ which proves (c). Furthermore, $A^\circ = (\bar{A})^\circ$ and therefore, by (b) (replacing A with $\bar{A}$),

$$\bar{A} \ominus G = (\bar{A})^\circ \ominus G = A^\circ \ominus G,$$

whence relation (d) follows because

$$A^\circ \ominus G \subset A \ominus G \subset \bar{A} \ominus G.$$

By (c), $A \ominus \lambda G = A \ominus \lambda \operatorname{co} G$ for all $\lambda \in \mathbb{R}$ and, since $\lambda_1 \operatorname{co} G \subset \lambda_2 \operatorname{co} G$ for $0 \leqq \lambda_1 \leqq \lambda_2$, we have

$$A \ominus \lambda_2 G = A \ominus \lambda_2 \operatorname{co} G \subset A \ominus \lambda_1 \operatorname{co} G = A \ominus \lambda_1 G,$$

which proves (e).

Finally, let the assumptions of (f) be satisfied, $z \in (A + B) \ominus B$, and l belong to $\mathscr{X}^*$, the set of continuous linear functionals on $\mathscr{X}$. Then $z + B \subset A + B$; hence

$$lz + \inf lB \geqq \inf l(A + B) = \inf lA + \inf lB,$$

implying $lz \geqq \inf lA$. Since $l \in \mathscr{X}^*$ is arbitrary, it follows from the convex separation theorem that $z \in A$. Thus $(A + B) \ominus B \subset A$. Conversely, if $a \in A$ then $a + B \subset A + B$ so that $A \subset (A + B) \ominus B$. This proves (f). □

LEMMA 4.2. *Let X be a closed subset of $\mathbb{R}^k$, $\mathscr{T} \subset \mathbb{R}^k$, $\mathscr{T}$ convex and compact, G an open neighborhood of 0 in $\mathscr{X}$, $\alpha_0 > 0$, $\psi = (\psi_1, \psi_2) : X \to \mathbb{R}^m \times \mathscr{X}$ continuous, $d \in \mathscr{X}$, and $C \subset \mathscr{X}$ convex. Assume that*

$$\lim_j a_j = \lim_j \beta_j = \lim_j \gamma_j = 0, \quad \lim_j s_j = 1, \quad \alpha \in [0, \alpha_0),$$

$$0 \leqq \alpha + \beta_j \leqq \alpha_0 \quad \forall j = 1, 2, \cdots,$$

and

(1) $p \in \alpha_j \bar{B}_m + \{\psi_1(x) \mid x \in X \cap (\alpha + \beta_j)\mathscr{T}, \psi_2(x) + s_j G \subset C + \gamma_j(C - d)\} \quad \forall j = 1, 2, \cdots.$

Then

$$p \in \{\psi_1(x) \mid x \in X \cap \alpha\mathscr{T}, \psi_2(x) \in C \ominus G\}.$$

Proof. By (1), for every $j \in \{1, 2, \cdots\}$ and $g \in G$, there exist

$$x_j \in X, \quad z_j \in \bar{B}_m, \quad c_{g,j} \in C$$

such that

(2) $$p = \alpha_j z_j + \psi_1(x_j), \quad x_j \in (\alpha + \beta_j)\mathscr{T}, \quad \psi_2(x_j) + s_j g = (1 + \gamma_j)c_{g,j} - \gamma_j d.$$

We may assume that

$$\lim_j x_j = x_0 \in X \cap [0, \alpha_0]\mathscr{T}, \qquad \lim_j z_j = z \in \bar{B}_m,$$

otherwise choosing appropriate subsequences. Then (2) implies that $p = \psi_1(x_0)$, $x_0 \in \alpha\mathscr{T}$, and

$$\lim_j c_{g,j} = \lim_j (1 + \gamma_j)^{-1}[\psi_2(x_j) + s_j g + \gamma_j d] = \psi_2(x_0) + g \quad \forall g \in G.$$

Thus $\psi_2(x_0) + g \in \bar{C}$ for all $g \in G$, hence $\psi_2(x_0) \in \bar{C} \ominus G$ and, by Lemma 4.1, $\psi_2(x_0) \in C \ominus G$. □

LEMMA 4.3. *Let X be a closed subset of $\mathbb{R}^k$, $0 \in X$, $\mathcal{T} \subset \mathbb{R}^k$, $\mathcal{T}$ convex and compact, G an open neighborhood of 0 in $\mathcal{Z}$,*

$$\beta > 0, \quad m \in \{1, 2, \cdots, k\}, \quad w \in \mathbb{R}^m, \quad r, r_1 > 0, \quad \eta \in \mathcal{T}, \quad \gamma \in (0, 1),$$

and $\psi = (\psi_1, \psi_2) : X \to \mathbb{R}^m \times \mathcal{Z}$ a continuous function such that

(a) *$\psi'(0)$ exists and $\psi'(0)\eta \in \{w\} \times [C \ominus G - \psi_2(0)]$;*

(b) *$\psi_1'(0)\bar{B}_k \supset \beta \bar{B}_m$;*

(c) *$\eta + r\bar{B}_k \subset \mathcal{T}$, $[0, r_1][\eta + r\bar{B}_k] \subset X$, $r\psi_2'(0)\bar{B}_k \subset \gamma G$, $\psi_2(0) \in C$.*

Then for every $\varepsilon \in (0, r)$ there exists $\hat{\alpha} > 0$ such that

$$(*) \qquad \psi_1(0) + \alpha[w + (r - \varepsilon)\beta \bar{B}_m] \subset \{\psi_1(x) \mid x \in X \cap \alpha\mathcal{T}, \psi_2(x) + \alpha(1 - \gamma)G \subset C\} \quad \forall \alpha \in [0, \hat{\alpha}].$$

If, furthermore,

(d) *$\varphi : Q \to \mathbb{R}^m \times \mathcal{Z}$, $\hat{q} : X \to Q$ and $u_n : X \to \mathcal{U}$ for $n = 1, 2, \cdots$ satisfy Condition* 2.0 *and $\psi(x) = \varphi(\hat{q}(x))$ then*

$$\begin{aligned} &\varphi_1(\hat{q}(0)) + \alpha[w + (r - \varepsilon)\beta \bar{B}_m] \\ (**) \qquad &\subset \{\varphi_1(u_n(x)) \mid x \in X \cap \alpha\mathcal{T}, \varphi_2(u_n(x)) + \alpha(1 - \gamma)G \subset C, n \in \{1, 2, \cdots\}\} \\ &\subset \{\varphi_1(u) \mid u \in \mathcal{U}, \varphi_2(u) + \alpha(1 - \gamma)G \subset C\} \quad \forall \alpha \in (0, \hat{\alpha}]. \end{aligned}$$

Proof. Let $e_1, \cdots, e_m$ be the columns of the unit $m \times m$ matrix and $b_1, \cdots, b_m \in \bar{B}_k$ such that

$$\psi_1'(0)b_\mu = \beta e_\mu \quad \forall \mu = 1, \cdots, m.$$

Let

$$a(\theta) = \eta + \sum_\mu \theta_\mu b_\mu \quad \forall \theta = (\theta_1, \cdots, \theta_m) \in r\bar{B}_m.$$

Since $r\psi_2'(0)\bar{B}_k$ is compact and γG open, it follows from (c) that there exists $\gamma_1 \in (0, \gamma)$ such that

$$(1) \qquad a(\theta) \in \mathcal{T}, \quad [0, r_1]a(\theta) \in X, \quad \sum_\mu \theta_\mu \psi_2'(0)b_\mu \in \gamma_1 G \quad \forall \theta \in r\bar{B}_m.$$

Furthermore, by (a),

$$(2) \qquad \psi_1'(0)a(\theta) = \psi_1'(0)\eta + \sum_\mu \theta_\mu \psi_1'(0)b_\mu = w + \beta\theta.$$

Now assume that condition (d) is satisfied. For $\theta \in r\bar{B}_m$ and $\alpha \subset (0, r_1]$, we set

$$(3) \qquad \begin{aligned} h^n(\theta, \alpha) &= \frac{1}{\alpha}[\varphi(u_n(\alpha a(\theta))) - \psi(\alpha a(\theta))] \\ d(\theta, \alpha) &= \frac{1}{\alpha}[\psi(\alpha a(\theta)) - \psi(0) - \alpha\psi'(0)a(\theta)]. \end{aligned}$$

Let $0 < \varepsilon < r$. We observe that we can determine $\alpha_1 \in (0, r_1]$ sufficiently small so that

$$(4) \qquad |d_1(\theta, \alpha)| \leqq \tfrac{1}{2}\varepsilon\beta, \quad d_2(\theta, \alpha) \in \tfrac{1}{2}(\gamma - \gamma_1)G \quad \forall \alpha \in (0, \alpha_1], \theta \in r\bar{B}_m.$$

Since $\lim_n \varphi(u_n(\alpha a(\theta))) = \psi(\alpha a(\theta))$ uniformly for all θ and α, for each $\alpha \in (0, \alpha_0]$ we can determine $N(\alpha)$ such that

$$(5) \qquad |h_1^{N(\alpha)}(\theta, \alpha)| \leqq \tfrac{1}{2}\varepsilon\beta, \qquad h_2^{N(\alpha)}(\theta, \alpha) \in \tfrac{1}{2}(\gamma - \gamma_1)G.$$

For arbitrary $z \in (r-\varepsilon)\beta \bar{B}_m$ and $\alpha \in (0, \alpha_1]$, we consider the equation

$$\varphi_1(u_{N(\alpha)}(\alpha a(\theta)) = \varphi_1(\hat{q}(0)) + \alpha(w+z) = \psi_1(0) + \alpha(w+z). \tag{6}$$

We have

$$\begin{aligned} \varphi(u_{N(\alpha)}(\alpha a(\theta)) &= \psi(\alpha a(\theta)) + \alpha h^{N(\alpha)}(\theta, \alpha) \\ &= \psi(0) + \alpha[\psi'(0)a(\theta) + d(\theta, \alpha) + h^{N(\alpha)}(\theta, \alpha)], \end{aligned} \tag{7}$$

and thus, by (2), (6) is equivalent to

$$\theta = \frac{1}{\beta}[z - d_1(\theta, \alpha) - h_1^{N(\alpha)}(\theta, \alpha)].$$

By (3)-(5), the right side of the above equation defines a continuous mapping of $r\bar{B}_m$ into itself. Therefore, this equation admits a solution $\bar{\theta}$ that also satisfies (6).

Now let $\hat{\alpha} := \min(\alpha_1, \varepsilon/\gamma)$. Then, by (a), (c), (1), (4), (5), (7), and Lemma 4.1,

$$\begin{aligned} \varphi_2(u_{N(\alpha)}(\theta, \alpha)) &= \psi_2(0) + \alpha\left[\psi_2'(0)\eta + \sum_{\mu} \theta_\mu \psi_2'(0) b_\mu + d_2(\theta, \alpha) + h_2^{N(\alpha)}(\theta, \alpha)\right] \\ &\in \psi_2(0) + \alpha[C \ominus G - \psi_2(0)] + \gamma_1 \alpha G + \tfrac{1}{2}\alpha(\gamma - \gamma_1)G + \tfrac{1}{2}\alpha(\gamma - \gamma_1)G \\ &\subset \psi_2(0) + \alpha[C \ominus (1-\gamma)G - \psi_2(0)] \subset C \ominus \alpha(1-\gamma)G \quad \forall \alpha \in (0, \bar{\alpha}]. \end{aligned} \tag{8}$$

Thus, the first two relations of (1), (6) (with its solution $\bar{\theta}$), and (8) yield relation (**).

The same argument, in a simpler form, will prove the validity of relation (*) without the use of assumption (d) if we redefine Q as X, φ as ψ, $\hat{q}(x)$ as x, and $u_n(x)$ as x. □

LEMMA 4.4. *Let X be a closed subset of $\mathbb{R}^k$, $\mathcal{T} \subset \mathbb{R}^k$, $\mathcal{T}$ convex and compact, G an open neighborhood of 0 in $\mathcal{X}$,*

$$m \in \{1, 2, \cdots, k\}, \quad w \in \mathbb{R}^m, \quad r, r_1 > 0, \quad \beta > 0, \quad \eta \in \mathcal{T},$$
$$\alpha_0 > 0, \quad \bar{\alpha} \in [0, \alpha_0), \quad \bar{x} \in X \cap \bar{\alpha}\mathcal{T}, \quad \gamma \in (0, 1),$$

and $\psi = (\psi_1, \psi_2) : X \to \mathbb{R}^m \times \mathcal{X}$ continuous. Let, furthermore, $\varepsilon_1 \geqq 0$, $\tau \in [0, 1)$, I be the unit $m \times m$ matrix, and M the $m \times k$ matrix of the form $[I, 0]$. Assume that

(a) *$\psi'(\bar{x})$ exists and $\psi'(\bar{x})\eta \in \{w\} \times [C \ominus G - \psi_2(\bar{x})]$;*

(b) *$[\psi_1'(\bar{x}) + \varepsilon_1 M]\bar{B}_k \supset \beta \bar{B}_m$;*

(c) *$\eta + r\bar{B}_k \subset \mathcal{T}$, $\bar{x} + [0, r_1][\eta + r\bar{B}_k] \subset X$,*

$$r\psi_2'(\bar{x})\bar{B}_k \subset (1-\tau)\gamma G, \qquad \psi_2(\bar{x}) \in C \ominus \tau G.$$

Then for every $\varepsilon \in (0, r)$ there exists $\hat{\alpha} \in (0, \alpha_0 - \bar{\alpha})$ such that

$$\begin{aligned} &\psi_1(\bar{x}) + (\alpha - \bar{\alpha})[w + (r-\varepsilon)\beta \bar{B}_m] \\ &\quad \subset \{\psi_1(x) + \varepsilon_1 M[x - \bar{x} - (\alpha - \bar{\alpha})\eta] \mid x \in X \cap \alpha\mathcal{T}, \\ &\qquad \psi_2(x) + (\alpha - \bar{\alpha})(1-\gamma)(1-\tau)G \subset C \ominus \tau G\} \qquad \forall \alpha \in [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}]. \end{aligned}$$

In particular, if $z \in r\beta B_m$ then there exists a function $\alpha \to x_\alpha$ on $[\bar{\alpha}, \bar{\alpha} + \hat{\alpha}]$ such that

$$x_\alpha \in \alpha\mathcal{T}, \qquad x_\alpha - \bar{x} \in (\alpha - \bar{\alpha})\mathcal{T},$$
$$\psi_1(x_\alpha) = \psi_1(\bar{x}) + (\alpha - \bar{\alpha})(w + z) - \varepsilon_1 M[x_\alpha - \bar{x} - (\alpha - \bar{\alpha})\eta],$$
$$\psi_2(x_\alpha) + (\alpha - \bar{\alpha})(1-\gamma)(1-\tau)G \subset C \ominus \tau G.$$

Proof. Let $\alpha_0^\# = \alpha_0 - \bar{\alpha}$, $X^\# = X - \bar{x}$ and, for all $x \in X^\#$,

$$\psi_1^\#(x) = \psi_1(\bar{x} + x) + \varepsilon_1 Mx, \quad \psi_2^\#(x) = \psi_2(\bar{x} + x), \quad \psi^\#(x) = (\psi_1^\#, \psi_2^\#)(x)$$
$$G^\# = (1-\tau)G, \quad C^\# = C \ominus \tau G, \quad w^\# = w + \varepsilon_1 M\eta.$$

Then

$$\psi_1^{\#\prime}(0)\eta = \psi_1'(\bar{x})\eta + \varepsilon_1 M\eta = w + \varepsilon_1 M\eta = w^{\#}, \qquad \psi_1^{\#\prime}(0)\bar{B}_k \supset \beta\bar{B}_m$$

$$\psi_2^{\#}(0) = \psi_2(\bar{x}) \in C \ominus \tau G_2 = C^{\#}, \qquad r\psi_2^{\#\prime}(0)\bar{B}_k \subset \gamma G^{\#}$$

and, by Lemma 4.1,

$$\psi_2^{\#\prime}(0)\eta = \psi_2'(\bar{x})\eta \in C \ominus G - \psi_2(\bar{x})$$
$$\subset C^{\#} \ominus (1-\tau)G - \psi_2^{\#}(0) = C^{\#} \ominus G^{\#} - \psi_2^{\#}(0).$$

Thus the function $\psi^{\#}$ satisfies the assumptions of Lemma 4.3, with the symbols denoted by $^{\#}$ replacing similarly named symbols in Lemma 4.3. It follows then, by relation (*) of Lemma 4.3, that for every $\varepsilon \in (0, r)$ there exists $\hat{\alpha}$ such that

$$\psi_1(\bar{x}) + \alpha[w + (r-\varepsilon)\beta\bar{B}_m] + \alpha\varepsilon_1 M\eta = \psi_1^{\#}(0) + \alpha[w^{\#} + (r-\varepsilon)\bar{B}_m]$$
$$\subset \{\psi_1^{\#}(x) \,|\, x \in X^{\#} \cap \alpha\mathcal{T},\ \psi_2^{\#}(x) + \alpha(1-\gamma)G^{\#} \subset C^{\#}\}$$
$$= \{\psi_1(\bar{x}+x) + \varepsilon_1 Mx \,|\, x \in X^{\#} \cap \alpha\mathcal{T},\ \psi_2(\bar{x}+x) + \alpha(1-\gamma)(1-\tau)G \in C \ominus \tau G\}$$
$$\forall \alpha \in [0, \hat{\alpha}].$$

We observe that, since $\mathcal{T}$ is convex and $\bar{x} \in \bar{\alpha}\mathcal{T}$, the relation $x \in \alpha\mathcal{T}$ implies $\bar{x} + x \in (\bar{\alpha} + \alpha)\mathcal{T}$. If we replace $\bar{\alpha} + \alpha$ by α and, in the expression on the right-hand side, $\bar{x} + x$ by x, then the conclusions follow directly. □

Proof of Theorem 2.1. *Step* 1. We may, and shall, assume that $k \geqq m$. Indeed, let $r(x)$ be as described in condition (c). Then

$$\sigma := \sup\{|r(x)| \,|\, x \in X \cap [0, \alpha_0]\mathcal{T}\} < \infty$$

because $\mathcal{T}$ is compact. Therefore, if $k < m$, we may replace

$$k,\ X,\ \mathcal{T},\ \psi(x_1, \cdots, x_k), \qquad \eta(x_1, \cdots, x_k)$$

by

$$k^{\#} = m, \quad X^{\#} = X \times \mathbb{R}^{m-k}, \quad \mathcal{T}^{\#} = \mathcal{T} \times \sigma\bar{B}_{m-k},$$

$$\psi^{\#}(x_1, \cdots, x_m) = \psi(x_1, \cdots, x_k), \qquad \eta^{\#}(x_1, \cdots, x_m) = (\eta(x_1, \cdots, x_k), 0, \cdots, 0),$$

which yields an equivalent problem that also satisfies conditions (a)-(c). (It is clear that, for $k < m$, conditions (e) and (f) cannot be satisfied.)

Let $s \in (0, 1)$, $T = \max\{|x| \,|\, x \in \mathcal{T}\}$ and

$$P_{\alpha,s} = \{\psi_1(x) \,|\, x \in X \cap \alpha\mathcal{T},\ \psi_2(x) \in C \ominus (1 - e^{-\alpha})sG\} \quad \forall \alpha \in [0, \alpha_0].$$

We shall show that, for all $\varepsilon \in (0, 1]$ and $\alpha \in [0, \alpha_0]$,

$$d[\psi_1(0) + \alpha w, P_{\alpha,s}] \leqq 2\varepsilon\alpha T. \tag{1}$$

With s and ε fixed, let

$$A = \{a \in [0, \alpha_0] \,|\, (1) \text{ is valid for } \alpha \in [0, \alpha]\}.$$

Each of the sets $P_{\alpha,s}$ and A is closed because ψ is continuous, $\mathcal{T}$ compact and, by Lemma 4.1, the set $C \ominus (1 - e^{-\alpha})sG$ closed. We easily verify that $0 \in A$. Now assume, by way of contradiction, that $\bar{\alpha} := \max A < a_0$. Since $\bar{\alpha} \in A$, there exists $\bar{x} \in X \cap \bar{\alpha}\mathcal{T}$ such that

$$|\psi_1(\bar{x}) - [\psi_1(0) + \bar{\alpha}w]| \leqq 2\varepsilon\bar{\alpha}T, \qquad \psi_2(\bar{x}) \in C \ominus (1 - e^{-\bar{\alpha}})sG. \tag{2}$$

Let M be the $m \times k$ matrix of the form $[I, 0]$, where I is the $m \times m$ unit matrix. We observe that $\psi_1'(\bar{x}) + \varepsilon_1 M$ is of full rank except for at most m values of ε_1. We may, therefore, choose $\varepsilon_1 \in (0, \varepsilon]$ so as to preserve full rank, and then determine $\beta > 0$ such that $[\psi_1'(\bar{x}) + \varepsilon_1 M]\bar{B}_k \supset \beta \bar{B}_m$. Let

$$\tau = s(1 - e^{-\bar{\alpha}}), \quad \gamma = (1 - s + se^{-\bar{\alpha}})^{-1}(1 - s), \quad r_1 = r_1(\bar{x}).$$

Since $\psi'(\bar{x})\bar{B}_k$ is compact, it follows from (c) that there exists $r \in (0, r(\bar{x})]$ such that

$$\eta(\bar{x}) + r\bar{B}_m \in \mathcal{T}, \quad \bar{x} + [0, r_1](\eta(\bar{x}) + r\bar{B}_m) \subset X, \quad r\psi_2'(\bar{x})\bar{B}_k \subset \gamma(1 - \tau)G.$$

It follows then, by Lemma 4.4 (with $\eta = \eta(\bar{x})$ and $z = 0$), that there exist $\hat{\alpha} \in (0, \alpha_0 - \bar{\alpha})$ and a function $\alpha \to x_\alpha : [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}] \to X$ such that, for all $\alpha \in [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}]$,

$$x_\alpha \in \alpha\mathcal{T}, \qquad x_\alpha - \bar{x} \in (\alpha - \bar{\alpha})\mathcal{T},$$

$$\psi_1(x_\alpha) = \psi_1(\bar{x}) + (\alpha - \bar{\alpha})w - \varepsilon_1 M[x_\alpha - \bar{x} - [\alpha - \bar{\alpha}]\eta(\bar{x})], \tag{3}$$

$$\psi_2(x_\alpha) + (\alpha - \bar{\alpha})(1 - \gamma)(1 - \tau)G \subset C \ominus \tau G.$$

This last relation implies (using Lemma 4.1) that

$$\psi_2(x_\alpha) \in C \ominus \sigma G,$$

where

$$\sigma = \tau + (\alpha - \bar{\alpha})(1 - \gamma)(1 - \tau),$$

and we verify that $\sigma \geqq s(1 - e^{-\alpha})$ for $\alpha \geqq \bar{\alpha}$. Thus (again by Lemma 4.1),

$$\psi_2(x_\alpha) \in C \ominus s(1 - e^{-\alpha})G \quad \forall \alpha \in [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}]. \tag{4}$$

Relations (2) and (3) imply that

$$\begin{aligned} |\psi_1(x_\alpha) - [\psi_1(0) + \alpha w]| &\leqq |\psi_1(x_\alpha) - \psi_1(\bar{x}) - (\alpha - \bar{\alpha})w| + |\psi_1(\bar{x}) - [\psi_1(0) + \bar{\alpha}w]| \\ &\leqq \varepsilon_1 |M[x_\alpha - \bar{x} - (\alpha - \bar{\alpha})\eta(\bar{x})]| + 2\varepsilon\bar{\alpha}T \leqq 2\varepsilon(\alpha - \bar{\alpha})T + 2\varepsilon\bar{\alpha}T \\ &= 2\varepsilon\bar{\alpha}T \qquad \forall \alpha \in [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}]. \end{aligned}$$

This last inequality, together with (4), shows that relation (1) is valid for all $\alpha \in [0, \bar{\alpha} + \hat{\alpha}]$, contrary to the definition of $\bar{\alpha}$. Thus (1) holds for all $\alpha \in [0, \alpha_0]$.

Since $P_{\alpha,s}$ is closed for each α and s and since ε can be chosen arbitrarily small, it follows that $\psi_1(0) + \alpha w \in P_{\alpha,s}$ for all $\alpha \in [0, \alpha_0]$ and $s \in (0, 1)$. By Lemma 4.2, this implies that validity of relation (*).

Step 2. We next proceed to prove relation (***) under assumptions (d) and (f). However, we at first replace (f) by the stronger assumption

H_1: assumption (f) holds with continuous $\rho(\cdot)$, $\beta(\cdot)$, and $\delta(\cdot)$.

Under assumptions (d) and H_1 we first prove that, for every

$$p \in B_m, \quad s \in (0, 1) \quad \text{and} \quad \alpha \in [0, \alpha_0],$$

we have

$$\psi_1(0) + \alpha w + sI(\alpha)p \in \{\psi_1(x) \mid x \in X \cap \alpha\mathcal{T}, \psi_2(x) \in C \ominus sJ(\alpha)G_2\} \tag{5}$$

for all $\alpha \in [0, \alpha_0]$. Let p and s be fixed, and let

$$A = \{a \in [0, \alpha_0] \mid (5) \text{ is valid for } \alpha \in [0, a]\}.$$

The functions $\alpha \to I(\alpha)$ and $\alpha \to J(\alpha)$ are continuous and positive and $J(\alpha) < 1$ for all $\alpha \in [0, \alpha_0]$. It follows that A is closed, and clearly $0 \in A$. Assume, by way of contradiction, that $\bar{\alpha} := \max A < \alpha_0$. Since $\bar{\alpha} \in A$, there exists $\bar{x} \in X \cap \bar{\alpha}\mathcal{T}$ such that

$$\psi_1(\bar{x}) = \psi_1(0) + \bar{\alpha}w + sI(\bar{\alpha})p, \qquad \psi_2(\bar{x}) \in C \ominus sJ(\bar{\alpha})G_2. \tag{6}$$

It follows now from (6) and Lemma 4.4, with the redefined parameters

$$\tau = sJ(\bar{\alpha}), \quad \varepsilon_1 = 0, \quad r = \rho(\bar{\alpha}), \quad r_1 = r_1(\bar{x}), \quad \beta = \beta(\bar{\alpha}),$$
$$\gamma = (1-\tau)^{-1}(1-\delta(\bar{\alpha})), \quad \varepsilon = r(1-|p|), \quad \eta = \eta(\bar{x})$$

that there exists $\hat{\alpha} \in (0, \alpha_0 - \bar{\alpha}]$ such that

$$\psi_1(\bar{x}) + (\alpha - \bar{\alpha})(w + r\beta[0,1]p) = \psi_1(0) + \alpha w + (sI(\bar{\alpha}) + (\alpha - \bar{\alpha})r\beta[0,1][0,1])p$$
$$\subset \{\psi_1(x) \,|\, x \in X \cap \alpha\mathcal{T}, \psi_2(x) \in C \ominus [\tau + (\alpha - \bar{\alpha})(\delta(\bar{\alpha}) - \tau)]G\} \tag{7}$$
$$\forall \alpha \in [\bar{\alpha}, \bar{\alpha} + \hat{\alpha}].$$

Since $\rho(\cdot)$, $\beta(\cdot)$, and $\delta(\cdot)$ are assumed to be positive and continuous and $0 < s < 1$, there exists $a_1 \in (0, \hat{\alpha}]$ such that, for all $\alpha \in [\bar{\alpha}, \bar{\alpha} + a_1]$,

$$(\alpha - \bar{\alpha})\rho(\bar{\alpha})\beta(\bar{\alpha}) \geqq s \int_{\bar{\alpha}}^{\alpha} \rho(t)\beta(t)\, dt = s[I(\alpha) - I(\bar{\alpha})],$$

and, therefore, there exists $\theta(\alpha) \in [0, 1]$ such that

$$sI(\bar{\alpha}) + (\alpha - \bar{\alpha})\rho(\bar{\alpha})\beta(\bar{\alpha})\theta(\alpha) = sI(\alpha). \tag{8}$$

Furthermore, we observe that $\alpha \to J(\alpha)$ satisfies the equation

$$J'(\alpha) + J(\alpha) = \delta(\alpha),$$

and, therefore, we can reduce a_1, if necessary, so that

$$\Delta J(\alpha)/\Delta\alpha + J(\bar{\alpha}) \leqq \frac{1}{s}\delta(\bar{\alpha}) \quad \text{for} \quad \alpha \in (\bar{\alpha}, \bar{\alpha} + a_1],$$

where

$$\Delta J(\alpha)/\Delta\alpha := (\alpha - \bar{\alpha})^{-1}[J(\alpha) - J(\bar{\alpha})].$$

If we recall that, in this step, $\tau = sJ(\bar{\alpha})$, then the last inequality implies

$$sJ(\alpha) \leqq sJ(\bar{\alpha}) + (\alpha - \bar{\alpha})[\delta(\bar{\alpha}) - sJ(\bar{\alpha})] = \tau + (\alpha - \bar{\alpha})[\delta(\bar{\alpha}) - \tau].$$

This relation, (7), and (8) yield

$$\psi_1(0) + \alpha w + sI(\alpha)p \in \{\psi_1(x) \,|\, x \in X \cap \alpha\mathcal{T}, \psi_2(x) \in C \ominus sJ(\alpha)G\}$$
$$\forall \alpha \in (\bar{\alpha}, \bar{\alpha} + a_1].$$

This shows that the relation (5) holds for all $\alpha \in [0, \bar{\alpha} + a_1]$, contradicting the definition of $\bar{\alpha}$, and thus proves that relation (5) holds for all $\alpha \in [0, \alpha_0]$.

We now replace assumption H_1 with the weaker assumption (f). We observe that assumption (f) remains satisfied if we replace $\rho(\cdot)$, $\beta(\cdot)$, $\delta(\cdot)$ by the continuous functions

$$\rho_n(\alpha) \leqq \rho(\alpha), \quad \beta_n(\alpha) \leqq \beta(\alpha), \quad \delta_n(\alpha) \leqq \delta(\alpha) \quad \forall j = 1, 2, \cdots, \quad \alpha \in [0, \alpha_0]$$

whose existence is postulated in (f). Since these functions are positive and converge pointwise from below, we have

$$\lim_n I_n(\alpha) = I(\alpha), \quad \lim_n J_n(\alpha) = J(\alpha) \quad \forall \alpha \in [0, \alpha_0],$$

where

$$I_n(\alpha) := \int_0^\alpha \rho_n(t)\beta_n(t)\,dt, \qquad J_n(\alpha) := \int_0^\alpha e^{-(\alpha-t)}\delta_n(t)\,dt,$$

and we have just shown that relation (5) is valid with $I(\alpha)$, $J(\alpha)$ replaced by $I_n(\alpha)$, $J_n(\alpha)$. We now choose a point $\bar{p} \in \bar{B}_m$ and sequences (s_n) in $(0,1)$ and (p_n) in B_m converging to 1 and $\bar{p}$. Since relation (5) holds with s, p, $I(\alpha)$, $J(\alpha)$ replaced by s_n, p_n, $I_n(\alpha)$, $J_n(\alpha)$, and since ψ is continuous and $\mathcal{T}$ compact, it follows from (5) (as modified) and Lemma 4.2 that

$$\psi_1(0) + \alpha w + I(\alpha)\bar{B}_m \subset \{\psi_1(x) \mid x \in X \cap \alpha\mathcal{T},\ \psi_2(x) + J(\alpha)G \subset C\} \quad \forall \alpha \in [0, \alpha_0]. \tag{9}$$

Thus relation (9) holds under assumptions (d) and (f). This proves relation (***).

Step 3. Our next goal is to prove relation (**) under assumptions (d) and (e). Let $0 < \bar{\alpha} < \alpha_0$, and let u_n be defined as in Condition 2.0. By (*) and (d), there exists $\bar{x} \in X \cap \bar{\alpha}\mathcal{T}$ such that

$$\psi_1(\bar{x}) = \varphi_1(\bar{q}) + \bar{\alpha}w, \qquad \psi_2(\bar{x}) = \varphi_2(\hat{q}(\bar{x})) \in C \ominus (1 - e^{-\bar{\alpha}})G.$$

Now fix $s \in (0,1)$. Then, by Condition 2.0, we may determine a sufficiently small $\rho \in (0, \alpha_0 - \bar{\alpha}]$ and a sufficiently large N so that

$$\psi_2(u_n(\bar{x}+x)) - \varphi_2(\hat{q}(\bar{x})) \in (1-s)(1-e^{-\bar{\alpha}})G;$$

hence, by Lemma 4.1,

$$\varphi_2(u_n(\bar{x}+x)) \in C \ominus (1-e^{-\bar{\alpha}})sG \quad \forall n \geqq N, \quad x \in X \cap [\bar{\alpha}, \bar{\alpha}+\rho]\mathcal{T}. \tag{10}$$

We than set

$$\begin{aligned} &\alpha_0^* = \rho, \quad \eta^* = 0 \in \mathbb{R}^k, \quad w^* = 0 \in \mathbb{R}^m, \quad X^* = X - x, \quad r^* = r(\bar{x}), \quad r_1^* = r_1(\bar{x}), \\ &\hat{q}^*(x) = \hat{q}(\bar{x}+x), \quad u_n^*(x) = u_n(\bar{x}+x), \quad \psi_1^*(x) = \varphi_1(\hat{q}^*(x)) \quad \forall x \in X^*. \end{aligned} \tag{11}$$

Since $\psi_1^{*\prime}(0) = \psi_1'(\bar{x})$ is of full rank, there exists $\beta^* > 0$ such that $\psi_1^{*\prime}(0)\bar{B}_k \supset \beta^*\bar{B}_m$. Thus the assumptions of Lemma 4.3 are satisfied by the function φ_1 and the objects defined in (11) (neglecting the relations in $\mathcal{X}$). Therefore, there exists $\hat{\alpha}^* \in (0, \alpha_0^*]$ (corresponding to $\varepsilon = 1 - s$) such that

$$\begin{aligned} \psi_1^*(0) + \alpha[w^* + s\beta^*\bar{B}_m] &= \varphi_1(\bar{q}) + \bar{\alpha}w + \alpha s\beta^*\bar{B}_m \\ &\subset \{\varphi_1(u_n^*(x)) \mid x \in X^* \cap \alpha\mathcal{T}\} \quad \forall \alpha \in (0, \hat{\alpha}^*]. \end{aligned}$$

Since $\mathcal{T}$ is convex, we have $\bar{x} + x \in (\bar{\alpha} + \alpha)\mathcal{T}$ and, therefore, in view of (10), this yields, for $\alpha = \alpha_1 = \min(\hat{\alpha}^*, \rho)$ and $V_{s,\bar{\alpha}} = \alpha_1 s\beta^*\bar{B}_m$,

$$\varphi_1(\bar{q}) + \bar{\alpha}w + V_{s,\bar{\alpha}} \subset \{\varphi_1(u) \mid u \in \mathcal{U},\ \varphi_2(u) + (1-e^{-\bar{\alpha}})sG \subset C\}.$$

This proves relation (**).

Step 4. We use a similar argument to prove relation (****). Specifically, assume that conditions (d) and (f) hold, and let $0 < \bar{\alpha} < \alpha_0$ and $p \in \bar{B}_m$. Then, as was shown in Step 2, relation (9) is valid and implies that there exists $\bar{x} \in X \cap \bar{\alpha}\mathcal{T}$ such that

$$\psi_1(\bar{x}) = \psi_1(0) + \bar{\alpha}w + I(\bar{\alpha})p, \qquad \psi_2(\bar{x}) = \varphi_2(\hat{q}(\bar{x})) \in C \ominus J(\bar{\alpha})G. \tag{12}$$

Now let $s \in (0, 1)$. We may determine a sufficiently small $\rho \in (0, a_0 - \bar{\alpha}]$ and a sufficiently large N so that

$$\varphi_2(u_n(\bar{x}+x)) \in C \ominus sJ(\bar{\alpha})G \quad \forall n \geqq N, \quad x \in X \cap (\bar{\alpha}+\rho)\mathscr{T}. \tag{13}$$

We then set

$$\begin{gathered} \alpha_0^* = \rho, \quad \eta^* = 0 \in \mathbb{R}^k, \quad w^* = 0 \in \mathbb{R}^m, \quad X^* = X - \bar{x}, \\ \beta^* = \beta(\bar{\alpha}), \quad \gamma^* = 1 - \delta(\bar{\alpha}), \quad r^* = r(\bar{x}), \quad r_1 = r_1(\bar{x}), \\ q^*(x) = \hat{q}(\bar{x}+x), \quad u_n^*(x) = u_n(\bar{x}+x), \quad \psi_1^*(x) = \varphi_1(\hat{q}^*(x)) \quad \forall x \in X^*. \end{gathered} \tag{14}$$

Then the assumptions of Lemma 4.3 are satisfied by the function φ_1 and the objects defined in (14) if we neglect the relations in $\mathscr{Z}$. Therefore, in view of (12), for every $\varepsilon \in (0, 1)$ there exists $\hat{\alpha} > 0$ such that

$$\begin{aligned} &\varphi_1(\bar{q}) + \bar{\alpha}w + I(\bar{\alpha})p = \psi_1(\bar{x}) = \varphi_1(\hat{q}(\bar{x})) \\ &\quad \in \varphi_1(\hat{q}^*(0)) + \alpha[w^* + (r^* - \varepsilon)\beta^* \bar{B}_m] \subset \{\varphi_1(u_n^*(x)) \,|\, x \in X^* \cap \alpha\mathscr{T},\ n = 1, 2, \cdots\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall \alpha \in (0, \hat{\alpha}]; \end{aligned}$$

hence, in view of relation (13) and setting $\alpha = \hat{\alpha}$ above, we have

$$\varphi_1(\bar{q}) + \bar{\alpha}w + I(\bar{\alpha})p \in \{\varphi_1(u) \,|\, u \in \mathscr{U},\ \psi_2(u) \in C \ominus sJ(\bar{\alpha})G\}.$$

This proves relation (****). □

Proof of Theorem 2.2. *Step* 1. Let $W = W(0)$ and ψ^i correspond to $\bar{x} = 0$, and let $s \in (0, 1)$. We can determine $\alpha_s > 0$ sufficiently small and i_s sufficiently large so that

$$X \cap \alpha_s\mathscr{T} \subset W, \quad \psi_2^i(x) - \psi_2(0) \in (1-s)G \quad \forall i \geqq i_s, \quad x \in X \cap \alpha_s\mathscr{T}.$$

Now let $i \geqq i_s$. Then ψ^i restricted to $X \cap \alpha_s\mathscr{T}$ satisfies the assumptions (a)-(c) of Theorem 2.1, with α_0 and C replaced by α_s and C_s, where C_s is the closure of $C + (1-s)$ co G. Therefore, by relation (*) of Theorem 2.1, for every $\alpha \in [0, \alpha_s]$ we have

$$\psi_1^i(0) + \alpha w \in \{\psi_1^i(x) \,|\, x \in X \cap \alpha\mathscr{T},\ \varphi_2^i(x) + (1 - e^{-\alpha})G \subset C_s\}$$

so that for each $i \geqq i_s$ there exists $x^i \in X \cap \alpha\mathscr{T}$ such that

$$\psi_1^i(0) + \alpha w = \psi_1^i(x^i), \qquad \psi_2^i(x^i) + (1 - e^{-\alpha})G \subset C_s.$$

We may choose a subsequence of (x^i) converging to some $x^\alpha \in X \cap \alpha\mathscr{T}$ and conclude that

$$\psi_1(0) + \alpha w = \psi_1(x^\alpha), \qquad \psi_2(x^\alpha) + (1 - e^{-\alpha})G \subset C_s,$$

which implies that, for all $\alpha \in [0, \alpha_s]$,

$$\psi_1(0) + \alpha w \in \{\psi_1(x) \,|\, x \in X \cap \alpha\mathscr{T},\ \psi_2(x) + (1 - e^{-\alpha})G \subset C_s\}. \tag{1}$$

Let

$$A = \{a \in [0, \alpha] \,|\, \text{relation (1) holds for all } \alpha \in [0, a]\}.$$

The set A is closed because ψ is continuous and $\mathscr{T}$ compact, and clearly $0 \in A$. Thus $\bar{\alpha} := \sup A \in A$ and there exists $\bar{x} \in X \cap \bar{\alpha}\mathscr{T}$ such that

$$\psi_1(\bar{x}) = \psi_1(0) + \bar{\alpha}w, \qquad \psi_2(\bar{x}) \in C_s \ominus (1 - e^{-\bar{\alpha}})G.$$

Now assume that $\bar{\alpha} < \alpha_0$, and set, for $X^* = X - \bar{x}$ and $x \in X^* \cap (\alpha_0 - \bar{\alpha})\mathscr{T}$,

$$\psi^*(x) = \psi(\bar{x}+x), \quad C^* = C_s \ominus (1 - e^{-\bar{\alpha}})G, \quad G^* = e^{-\bar{\alpha}}G.$$

By Lemma 4.1, C^* is convex and $C^* \ominus G^* = C_s \ominus G$. Therefore, our argument leading to relation (1) is valid, with ψ, α_0, C, and G replaced by ψ^*, $\alpha_0 - \bar{\alpha}$, C^*, and G^*, and implies that there exists $\alpha_s \in (0, \alpha_0 - \bar{\alpha}]$ such that

$$\psi_1^*(0) + \alpha w \in \{\psi_1^*(x) \mid x \in X^* \cap \alpha \mathcal{T}, \psi_2^*(x) + (1 - e^{-\alpha}) G^* \subset C^*\} \tag{2}$$

$$\forall \alpha \in [0, \alpha_s].$$

We have

$$1 - e^{-\bar{\alpha}} + e^{-\bar{\alpha}}(1 - e^{-\alpha}) = 1 - e^{-(\bar{\alpha}+\alpha)}$$

and therefore, by Lemma 4.1,

$$C^* \ominus (1 - e^{-\alpha}) G^* = C_s \ominus (1 - e^{-(\bar{\alpha}+\alpha)}) G.$$

Thus, by (2), for all $\alpha \in [0, \alpha_s]$ we have

$$\psi_1(0) + (\bar{\alpha} + \alpha) w \in \{\psi_1(x) \mid x \in X \cap (\bar{\alpha} + \alpha) \mathcal{T}, \psi_2(x) + (1 - e^{-(\bar{\alpha}+\alpha)}) G \subset C_s\}.$$

This shows that $\max A \geqq \bar{\alpha} + \alpha_s > \bar{\alpha}$, contradicting the definition of $\bar{\alpha}$. Therefore, $A = [0, \alpha_0]$ and relation (1) is valid for all $s \in (0, 1)$ and $\alpha \in [0, \alpha_0]$. Our final conclusion now follows from (1) and Lemmas 4.1 and 4.2 by letting $s \to 1$. □

REFERENCES

[1] F. CSAKI, *Modern Control Theories: Nonlinear, Optimal and Adaptive Systems*, Akademiai Kiado, Budapest, 1972.

[2] H. FRANKOWSKA, *The first order necessary conditions for nonsmooth variational and control problems*, SIAM J. Control Optim., 22 (1984), pp. 1-12.

[3] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[4] B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[5] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[6] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, SIAM J. Control Optim., 14 (1976), pp. 546-572.

[7] ———, *Fat homeomorphisms and unbounded derivative containers*, J. Math. Anal. Appl., 81 (1981), pp. 545-560; 90 (1982), pp. 582-583.

[8] ———, *Optimization and controllability without differentiability assumptions*, SIAM J. Control Optim., 21 (1983), pp. 837-855.

[9] ———, *Higher order conditions with and without Lagrange multipliers*, SIAM J. Control Optim., 24 (1986), pp. 715-730.

[10] ———, *Homeomorphisms and local C1 approximations*, J. Nonlinear Anal. TMA, 12 (1988), pp. 593-597.

[11] ———, *Higher order conditions for conical controllability*, SIAM J. Control Optim., 26 (1988), pp. 1471-1480.

[12] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, Berlin, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 991–1006, September 1989

006

OPTIMAL FEEDBACK CONTROLS*

LEONARD D. BERKOVITZ†

This paper is dedicated to the memory of E. J. McShane.

Abstract. Optimal control problems governed by ordinary differential equations with control constraints that are not necessarily compact are considered. Conditions imposed on the data and on the structure of the terminal sets imply that the minimum is attained and that the value function is locally Lipschitz. A necessary condition in terms of lower directional Dini derivates of the value function is given. The condition reduces to the Bellman-Hamilton-Jacobi (BHJ) condition at points of differentiability of the value, and for a subclass of the problems considered implies that the value is a viscosity solution of the BHJ equation. A strengthened version of the necessary condition gives an optimal feedback control and a procedure for approximating optimal controls.

Key words. optimal feedback control, synthesis of optimal control, lower directional Dini derivates, Bellman-Hamilton-Jacobi equation, viscosity solution

AMS(MOS) subject classifications. 49B05, 49C05

1. Introduction. An important optimal control problem is the following. The state of a system at time t is described by an n-vector $x(t)=(x^1(t), \cdots, x^n(t))$ whose evolution is governed by a system of differential equations

$$x' = f(t, x, u(t)), \qquad x(\tau) = \xi \tag{1.1}$$

where u is a control function, or simply control, selected from some preassigned class of functions. A terminal set $\mathcal{T}$ is given in (t, x)-space, as is a real-valued function g defined on $\mathcal{T}$. Let $\phi(\cdot) = \phi(\ , \tau, \xi, u)$ be a solution of (1.1) such that at some time t_f, the trajectory $\phi(\cdot)$ hits the set $\mathcal{T}$ for the first time. The optimal control problem is to select a control u^* that minimizes $g(t_f, \phi(t_f))$.

Another problem, whose solution is even more useful in applications, is that of finding the optimal feedback, or optimal control synthesis. In this problem we seek a function U in some class defined on a region $\mathcal{R}$ of (t, x) space such that for all initial points (τ, ξ) in $\mathcal{R}$ the solution of

$$x' = f(t, x, U(t, x)), \qquad x(\tau) = \xi \tag{1.2}$$

lies in $\mathcal{R}$ and minimizes $g(t_f, \phi(t_f))$ for the initial point selected.

Suppose that the values $u(t)$ of the controls u are required to lie in some set $Z(t)$ in $\mathbb{R}^n$. Suppose that for each (τ, ξ) in some region $\mathcal{R}$, the optimal control problem has a solution and that u^* and ϕ^* are the control and corresponding trajectory that achieve the minimum. Let $W(\tau, \xi)$ denote the value of the minimum for the problem with initial point (τ, ξ). If W is continuously differentiable on $\mathcal{R}$, if the data of the problem are sufficiently differentiable, and if u is piecewise continuous, then the well-known dynamic programming argument shows that all points (t, x) of $\mathcal{R}$

$$W_t(t, x) + \min_{z \in Z(t)} \langle W_x(t, x), f(t, x, z) \rangle = 0 \tag{1.3}$$

where $\langle\ ,\ \rangle$ denotes inner product. Moreover, the minimum is attained at the values of the optimal controls at time t for the problem with initial point (t, x).

* Received by the editors August 15, 1988; accepted for publication (in revised form) December 7, 1988. This research was supported by National Science Foundation Grant DMS 8700813.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

Relation (1.3) can also be used as a sufficient condition for optimality, an observation first made by Carathéodory [4, § 231] in connection with problems in the calculus of variations. Let W be a continuously differentiable function such that

$$W(t_f, x_f) = g(t_f, x_f) \tag{1.4}$$

at points (t_f, x_f) of the terminal set. Let u^* and ϕ^* be a control and corresponding trajectory for the problem with initial point (τ, ξ) such that if for all $\tau \leqq t \leqq t_f^*$, whenever we substitute $x = \phi^*(t)$ in (1.3), then the minimum is attained at $z = u^*(t)$. Then (ϕ^*, u^*) is optimal. To see this we note, assuming the requisite differentiability, that if u is a control and ϕ is the corresponding trajectory that

$$\begin{aligned} W(t_f, \phi(t_f)) - W(\tau, \xi) &= \int_\tau^{t_f} \left\{ \frac{d}{ds} W(s, \phi(s)) \right\} ds \\ &= \int_\tau^{t_f} \{ W_t(t, \phi(t)) + \langle W_x(t, \phi(t)), f(t, \phi(t), u(t)) \rangle \} \, ds. \end{aligned}$$

By (1.3) the integrand is nonnegative, so that $W(t_f, \phi(t_f)) \geqq W(\tau, \xi)$. On the other hand, if we take $(\phi, u) = (\phi^*, u^*)$, the integrand is zero and so $W(t_f^*, \phi^*(t_f^*)) = W(\tau, \xi)$. But by (1.4), $W(t_f, \phi(t_f)) = g(t_f, \phi(t_f))$ and $W(t_f^*, \phi^*(t_f^*)) = g(t_f^*, \phi(t_f^*))$. Thus (ϕ^*, u^*) minimizes.

Suppose further that a function U^* can be found such that the minimum in (1.3) is attained at $z = U^*(t, x)$ and such that (1.2) with $U(t, x)$ replaced by $U^*(t, x)$ has solutions for all (τ, ξ). If the function U^* has the requisite smoothness properties, the arguments of the preceding paragraph show that U^* is an optimal feedback control.

Unfortunately, in many specific problems the sufficiency theorem and the method of obtaining optimal feedback controls outlined above cannot be applied because the value function W does not have the requisite smoothness properties, nor does there exist any other solution W of (1.2) with the requisite smoothness.

Various authors have attempted to salvage the Carathéodory approach. One technique applicable to many examples is to assume that the field of extremals—that is, trajectories along which the Pontryagin maximum principle holds—has a certain structure. The ideas of Carathéodory can then be adapted to these situations. See Berkovitz [2], Boltyanskii [3], and Young [12], for example. Another approach, in which (1.3) is replaced by an equation involving the Clarke generalized gradient, was taken by Vinter and by Clarke and Vinter. See [6] and [7]. Reference [7] gives references to earlier work of Vinter.

In this paper we shall develop yet another modification of Carathéodory's ideas. We first recall a definition. Let L be a real-valued function defined on $\mathbb{R} \times \mathbb{R}^n$. The lower Dini derivate of L at the point (t, x) in the direction $(1, h)$, where $h \in \mathbb{R}^n$, is denoted by $D^- L(t, x; 1, h)$ and is defined by

$$D^- L(t, x; 1, h) = \liminf_{\delta \downarrow 0} \frac{L(t + \delta, x + \delta h) - L(t, x)}{\delta}. \tag{1.5}$$

Similarly, the upper Dini derivate of L at (t, x) in the direction $(1, h)$ is denoted by $D^+ L(t, x; 1, h)$ and is defined as in (1.5) with lim inf replaced by lim sup. The function L is said to have a directional derivative at (t, x) in the direction $(1, h)$ if $D^+ L(t, x; 1, h) = D^- L(t, x; 1, h)$. We denote the directional derivative by $DL(t, x; 1, h)$. If L is differentiable at (t, x), then $DL(t, x; 1, h)$ exists for every $h \in \mathbb{R}^n$ and

$$DL(t, x; 1, h) = L_t(t, x) + \langle L_x(t, x), h \rangle.$$

To describe our approach as simply as possible we shall assume that for each t the set $Z(t)$ is a fixed compact set Z. In the paper itself more general constraints will be allowed. Let $Q(t, x) = \{h: h = f(t, x, z), z \in Z\}$. In the problems that we consider the value function W will be locally Lipschitz continuous. We shall show that W satisfies

$$\min_{h \in Q(t,x)} D^- W(t, x; 1, h) = 0 \tag{1.6}$$

at each point (t, x) of $\mathcal{R}$. This is the generalization of (1.3). At points of differentiability of W, relation (1.6) implies (1.3). Relation (1.6) also implies that W is a viscosity solution of the Hamilton–Jacobi equation (1.3).

Now let W be any locally Lipschitz function that satisfies (1.6) at all points of $\mathcal{R}$ and that satisfies the boundary condition (1.4) at points of $\mathcal{T}$. For (t, x) in $\mathcal{R}$ let

$$\begin{aligned} F(t, x) &= \arg\min_{h \in Q(t,x)} D^- W(t, x; 1, h) \\ &\equiv \{h: h \in Q(t, x), D^- W(t, x; 1, h) = 0\}. \end{aligned} \tag{1.7}$$

Then $F(t, x)$ is nonempty, and the function F may be set-valued. It is an easy calculation, which we shall give in Lemma 5.1 below, that any absolutely continuous solution ψ of the differential inclusion

$$x' \in F(t, x), \qquad x(\tau) = \xi \tag{1.8}$$

is optimal. Thus, if (1.8) has a solution for each initial point (τ, ξ) in $\mathcal{R}$, then F will be an optimal feedback control in the sense described. In this paper we shall impose conditions on F that guarantee the existence of a solution of (1.8). The proof will describe a constructive method for approximating such solutions.

In conclusion we point out that with minor modifications in wording and hypotheses to suit the new context, our results and arguments apply to control problems in which the state of the system is governed by a differential inclusion

$$x' \in Q(t, x), \qquad x(\tau) = \xi.$$

The interested reader will have no trouble in interpreting our results in this context.

Results related to some of ours have been obtained in the differential inclusion context by Frankowska [8].

2. Assumptions and problem formulation. As indicated in the Introduction, t denotes time and x denotes a vector in $\mathbb{R}^n$, the state space. The letter z denotes a vector in $\mathbb{R}^m$, the range space of the controls. Components of vectors are denoted by superscripts. Thus $x = (x^1, \cdots, x^n)$, $z = (z^1, \cdots, z^m)$, etc. The Euclidean norm of a vector is denoted by $|\cdot|$. Let $T > 0$ be fixed.

A mapping from $[0, T] \times \mathbb{R}^n \times \mathbb{R}^m$ to $\mathbb{R}^n$ is denoted by f, and g denotes a mapping from $[0, T] \times \mathbb{R}^n$ to $\mathbb{R}$. The letter Z denotes a mapping from $[0, T]$ to the subsets of $\mathbb{R}^m$. If $\mathcal{Y}$ is a subset of some Euclidean space, by an ε-neighborhood of $\mathcal{Y}$, we mean the set $\mathcal{N}_\varepsilon(\mathcal{Y})$ defined by $\mathcal{N}_\varepsilon(\mathcal{Y}) = \{y': |y' - y| < \varepsilon \text{ for some } y \in \mathcal{Y}\}$.

A mapping Q from a subset $\mathcal{Y}$ of a Euclidean space $\mathbb{R}^p$ to subsets of a Euclidean space $\mathbb{R}^q$ is said to satisfy property (Q) at a point y in $\mathcal{Y}$ if

$$Q(y) = \bigcap_{\delta > 0} \operatorname{clco} \{Q(\mathcal{N}_\delta(y) \cap \mathcal{Y})\}. \tag{2.1}$$

Note that if the mapping Q satisfies property (Q) then $Q(y)$ must be closed and convex.

We now state our assumptions on the data of the problem.

ASSUMPTION I. (1) The mapping f is continuous on $[0, T]\times\mathbb{R}^n\times\mathbb{R}^m$.

(2) For every $R>0$, there exists a constant $K_R>0$ such that for all $t\in[0, T]$, $z\in\mathbb{R}^m$ and x, $\bar{x}$ in $\mathbb{R}^n$ with $|x|\leqq R$, $|\bar{x}|\leqq R$,

$$|f(t, x, z)-f(t, \bar{x}, z)|\leqq K_R|x-\bar{x}|.$$

(3) There exists a real-valued function β in $L_\infty[0, T]$ such that for almost all $t\in[0, T]$ and all $x\in\mathbb{R}^n$, $z\in\mathbb{R}^m$

$$|f(t, x, z)|\leqq\beta(t)(1+|x|).$$

(4) For each (t, x) in $[0, T]\times\mathbb{R}^n$ let

$$Q(t, x)=\{h: h=f(t, x, z),\ z\in Z(t)\}.$$

The mapping Q satisfies property (Q) at every point of $[0, T]\times\mathbb{R}^n$.

(5) For every $R>0$ there exists a constant $K'_R>0$ such that for all t, $\bar{t}$ in $[0, T]$ and all x, $\bar{x}$ such that $|x|\leqq R$, $|\bar{x}|\leqq R$,

$$|g(\bar{t}, \bar{x})-g(t, x)|\leqq K'_R(|\bar{t}-t|+|\bar{x}-x|).$$

Remark 2.1. The requirement that Q satisfies property (Q) is imposed to ensure the existence of optimal solutions, since we do not assume that the sets $Z(t)$ are compact. The mapping $Z=Z(\cdot)$ is said to be upper semicontinuous with respect to inclusion (u.s.c.i) at t_0 if for every $\varepsilon>0$ there exists a $\delta>0$ such that $Z(t)\subseteq\mathcal{N}_\varepsilon(Z(t_0))$ whenever $|t-t_0|<\delta$. In problems with compact constraint sets $Z(t)$, in order to guarantee the existence of optimal solutions it is assumed that at each t in $[0, T]$ the set $Z(t)$ is compact and the mapping Z is u.s.c.i. It is also assumed that at each (t, x) in $[0, T]\times\mathbb{R}^n$ the set $Q(t, x)$ is convex. It is known that for f continuous, these conditions imply that Q satisfies property (Q) in $[0, T]\times\mathbb{R}^n$. Thus, our treatment includes the compact case. See [1] and [5].

Remark 2.2. If the sets $Q(t, x)$ are not convex, we consider the relaxed problem in which the right-hand side of (1.1) is replaced by

$$\tilde{f}(t, x, \tilde{z}, \pi)\equiv\sum_{i=1}^{n+1}\pi^i f(t, x, z_i)$$

where $\tilde{z}=(z_1, \cdots, z_{n+1})\in\mathbb{R}^{(n+1)m}$ and $\pi=(\pi^1, \cdots, \pi^{n+1})\in\mathbb{R}^{n+1}$. The state variable for this problem is x and the control variables are $\tilde{z}$ and π. If $z\in Z(t)$ is the control constraint for the original problem, then the control constraint for the relaxed problem is $\tilde{z}\in\tilde{Z}(t)$, where $\tilde{Z}(t)$ denotes the Cartesian product of $Z(t)$ with itself $(n+1)$ times, and $\pi\in\Gamma\equiv\{\pi: \pi^i\geqq 0, \sum_{i=1}^{n+1}\pi^i=1\}$. In the relaxed problem the sets

$$\tilde{Q}(t, x)=\{h: h\in\tilde{f}(t, x, \tilde{z}, \pi),\ (\tilde{z}, \pi)\in\tilde{Z}(t)\times\Gamma\}$$

are convex. We henceforth suppose that if the sets $Q(t, x)$ are not convex, then we are considering the relaxed problem. We shall, however, retain the notation of the original problem. Note, however, that in the noncompact case we must still postulate that $\tilde{Q}(t, x)$ is closed and that (2.1) holds at each point $y=(t, x)$ in $[0, T]\times\mathbb{R}^n$.

A *control* u on $[\tau, T]$, $0\leqq\tau<T$, will be defined to be a measurable function defined on $[\tau, T]$ with $u(t)\in Z(t)$ almost everywhere on $[\tau, T]$.

Standard theorems concerning the existence and uniqueness of solutions of ordinary differential equations, elementary arguments and Gronwall's Lemma give the following result.

LEMMA 2.1. *Let f satisfy* (1)-(3) *of Assumption* I. *Then for each* (τ, ξ) *in* $[0, T)\times\mathbb{R}^n$ *and each control* u *on* $[\tau, T]$ *the differential equation* (1.1) *has a unique solution* $\phi(\ , \tau, \xi)$ *defined on* $[\tau, T]$. *Let* $\mathscr{K}$ *be a compact set contained in* $[0, T)\times\mathbb{R}^n$. *Then these exists a constant* $R>0$ *that depends only on* $\mathscr{K}$, *such that for any* (τ, ξ) *in* $\mathscr{K}$ *and any control* u *on* $[\tau, T]$ *the unique solution* $\phi(\ , \tau, \xi)$ *of* (1.1) *satisfies*

$$|\phi(t, \tau, \xi)| \leqq R \quad \text{for } \tau \leqq t \leqq T \tag{2.2}$$

and

$$|\phi(t, \tau, \xi) - \phi(t', \tau, \xi)| \leqq (1+R)|t - t'| \tag{2.3}$$

for all t, t' *in* $[\tau, T]$. *Moreover, there exists a constant* $K_1>0$ *that depends only on* $\mathscr{K}$ *such that for any pair of points* (τ, ξ) *and* (τ', ξ') *in* $\mathscr{K}$ *and any control* u *defined on* $[\max(t, \tau'), T]$

$$|\phi(t, \tau, \xi) - \phi(t, \tau', \xi')| \leqq K_1(|\tau - \tau'| + |\xi - \xi'|) \tag{2.4}$$

for $\max(\tau, \tau') \leqq t \leqq T$.

Given an initial point (τ, ξ) and a control u on $[\tau, T]$, we shall refer to $\phi(\ , \tau, \xi)$ as the trajectory corresponding to u.

We now discuss the nature of the terminal set $\mathscr{T}$. Let $\mathscr{F}$ be a closed domain in $[0, \infty)\times\mathbb{R}^n$ whose boundary is a $C^{(2)}$ manifold. We suppose that the intersection of $\mathscr{F}$ with the hyperplane $t = T$ has nonvoid interior relative to the hyperplane.

If $[T, \infty)\times\mathbb{R}^n \not\subseteq \mathscr{F}$, then we define

$$\mathscr{T} = \mathscr{F} \cup ([T, \infty)\times\mathbb{R}^n) \tag{2.5}$$

and say that the terminal set $\mathscr{T}$ is of Type I.

Example. To illustrate terminal sets of Type I consider the problem of reaching an ε-neighborhood of the origin in minimum time, where we restrict the time so as not to exceed some fixed time T. The set $\mathscr{F}$ for this problem is the half-cylinder $\{(t, x): t \geqq 0, |x| \leqq \varepsilon\}$ and the set $\mathscr{T}$ is the union of this half cylinder and the halfspace $[T, \infty)\times\mathbb{R}^n$. The function g is defined by $g(t, x) = t$.

If $[T, \infty)\times\mathbb{R}^n \subseteq \mathscr{F}$, then we define

$$\mathscr{T} = \mathscr{F} \tag{2.6}$$

and say that the terminal set $\mathscr{T}$ is of Type II.

Let

$$\mathscr{R} = \{(\tau, \xi): (\tau, \xi) \in [0, T)\times\mathbb{R}^n, (\tau, \xi) \notin \mathscr{T}\}. \tag{2.7}$$

It follows from Lemma 2.1 and the definition of $\mathscr{T}$, that for any (τ, ξ) in $\mathscr{R}$ and any control u on $[\tau, T]$, the corresponding trajectory $\phi(\ , \tau, \xi)$ will intersect $\mathscr{T}$ at a first time $t_1 = t_1(\tau, \xi, u) = \min\{t: (t, \phi(t, \tau, \xi)) \in \mathscr{T}\}$. We call t_1 the terminal time of the trajectory ϕ.

The stopping rule for our problem will be, "Stop at time t_1."

For each (τ, ξ) in $\mathscr{R}$ such that controls exist on $[\tau, T]$ we formulate the following problem.

PROBLEM I. Minimize $J(\tau, \xi, u) \equiv g(t_1, \phi(t_1, \tau, \xi))$ subject to (1.1) over all controls u on $[\tau, T]$.

It is known that under Assumption I and our assumptions on $\mathscr{T}$ that if the set of controls is nonempty, then there exists a control u^* on $[\tau, T]$ that furnishes a minimum. See [1] and [5].

To carry out our analysis we will need to impose a further condition relating the dynamics to the boundary. If $\mathcal{T}$ is of Type I, let $\hat{\partial}\mathcal{F}$ denote those boundary points of $\mathcal{F}$ that do not belong to $(T, \infty)\times\mathbb{R}^n$, i.e.,

$$\hat{\partial}\mathcal{F} = \partial\mathcal{F} \cap [0, T]\times\mathbb{R}^n. \tag{2.8}$$

If $\mathcal{T}$ is of Type II, let $\hat{\partial}\mathcal{F} = \partial\mathcal{T} = \partial\mathcal{F}$.

Note that $\hat{\partial}\mathcal{F}$ is smooth, since it is a subset of the boundary of $\mathcal{F}$. In our example $\hat{\partial}\mathcal{F} = \{(t, x)\colon 0 \leqq t \leqq T,\ |x| = \varepsilon\}$.

For $p^0 \in \mathbb{R}$, $p \in \mathbb{R}^n$, define

$$B(t, x, p_0, p) = \sup\{p_0 + \langle p, f(t, x, z)\rangle\colon z \in Z(t)\}. \tag{2.9}$$

At $(t, x) \in \hat{\partial}\mathcal{F}$ let $(\nu_0, \nu) = (\nu_0(t, x), \nu(t, x))$ denote the unit normal to $\partial\mathcal{F}$ that points to the exterior of $\mathcal{T}$.

ASSUMPTION II. For every compact subset $\mathcal{B}$ of $\hat{\partial}\mathcal{F}$ there exists an $\varepsilon > 0$ and a constant $c > 0$, where ε and c depend only on $\mathcal{B}$ such that the following holds. If $(t', x') \in \mathcal{B}$, if $(t, x) \in \mathcal{N}_\varepsilon(t', x')$ and if $(p^0, p) \in \mathcal{N}_\varepsilon(\nu_0(t', x'), \nu(t', x'))$ then $B(t, x, p_0, p) \leqq -c$.

Remark 2.3. If for each t, the set $Z(t)$ is compact and the mapping Z is u.s.c.i. on $[0, T]$, then the assumption

$$\nu^0 + \langle \nu, f(t, x, z)\rangle < 0$$

for all $(t, x) \in \hat{\partial}\mathcal{F}$ implies Assumption II.

3. Lipschitz continuity of the value. We henceforth assume that for each $0 \leqq \tau < T$ there exists at least one control on $[\tau, T]$. We have already noted that under this assumption Problem I has a solution for each initial point (τ, ξ) in $\mathcal{R}$. Let $W(\tau, \xi)$ denote the value of the minimum for the problem with initial point (τ, ξ). The function W is thus defined on all of $\mathcal{R}$.

We extend the definition of W to $\mathcal{R} \cup \partial\mathcal{T} = \bar{\mathcal{R}}$ by the formula

$$W(t_1, x_1) = g(t_1, x_1) \quad \text{if } (t_1, x_1) \in \partial\mathcal{T}. \tag{3.1}$$

We now state the principal result of § 3.

THEOREM 3.1. *Let Assumptions* I *and* II *hold and let* W *be defined as above. Then for every compact set* $\mathcal{K}$ *contained in* $\mathcal{R} \cup \partial\mathcal{T}$, *there exists a constant* $K > 0$ *such that for* (τ, ξ) *and* $(\bar{\tau}, \bar{\xi})$ *in* $\mathcal{K}$,

$$|W(\tau, \xi) - W(\bar{\tau}, \bar{\xi})| \leqq K(|\tau - \bar{\tau}| + |\xi - \bar{\xi}|). \tag{3.2}$$

COROLLARY 3.1. $W(\tau, \xi) \to g(t_1, x_1)$ *as* $(\tau, \xi) \to (t_1, x_1)$, *for all* $(t_1, x_1) \in \partial\mathcal{T}$, *and the convergence is uniform on compact subsets of* $\partial\mathcal{T}$.

The corollary is an immediate consequence of the theorem and (3.1).

Our proof will utilize the function ρ defined on $[0, \infty)\times\mathbb{R}^n$ as follows: $\rho(t, x) =$ signed distance of (t, x) to $\partial\mathcal{F}$, where we take $\rho(t, x) > 0$ if $(t, x) \notin \mathcal{F}$ and $\rho(t, x) < 0$ if $(t, x) \in (\mathcal{F} - \partial\mathcal{F})$. Since $\partial\mathcal{F}$ is $C^{(2)}$ it follows that if $\mathcal{B}$ is a compact subset of $\partial\mathcal{F}$, then there exists an $\varepsilon_0 > 0$ such that ρ is C' on $\mathcal{N}_{\varepsilon_0}(\mathcal{B})$. Also, at points (t_1, x_1) of $\partial\mathcal{F}$

$$(\rho_t(t, x), \rho_x(t, x)) \to (\nu^0, \nu)$$

as $(t, x) \notin \mathcal{F}$ tends to (t_1, x_1). Moreover, on compact subsets of $\partial\mathcal{F}$, the convergence is uniform. It therefore follows from Assumption II that for any compact subset $\mathcal{B}$ of $\partial\mathcal{F}$, and hence for any compact subset $\mathcal{B}$ of $\hat{\partial}\mathcal{F}$, there exist an $\varepsilon_1 > 0$ and a $c > 0$ such that for all $(t, x) \in \mathcal{N}_{\varepsilon_1}(\mathcal{B})$

$$\sup_{z \in Z(t)} [\rho_t(t, x) + \langle \rho_t(t, x), f(t, x, z)\rangle] \leqq -c. \tag{3.3}$$

Since $\mathscr{K}$ is compact it suffices to show that there exist a $K>0$ and a $\delta>0$, both of which depend only on $\mathscr{K}$, such that (3.2) holds whenever $|(\tau,\xi)-(\tau',\xi')|<\delta$.

We first consider the case in which both (τ,ξ) and $(\bar\tau,\bar\xi)$ are in $\mathscr{K}\cap\mathscr{R}$; neither point is on the boundary. To every control u on $[\tau,T]$ we associate a control $\bar u$ on $[\bar\tau,T]$ as follows. If $\bar\tau>\tau$, let $\bar u(t)=u(t)$ for $\bar\tau\leqq t\leqq T$. If $\bar\tau<\tau$, let $\hat u$ be any control on $[\bar\tau,T]$. Define $\bar u(t)=\hat u(t)$ for $\bar\tau\leqq t<\tau$ and $\bar u(t)=u(t)$ for $\tau\leqq t\leqq T$. Let $\phi(\ )=\phi(\ ,\tau,\xi)$ and $\bar\phi(\cdot)=\bar\phi(\ ,\bar\tau,\bar\xi)$ denote the trajectories corresponding to u and $\bar u$, respectively. Let $t_1=t_1(\tau,\xi)$ and $\bar t_1=\bar t_1(\bar\tau,\bar\xi)$ denote the corresponding terminal times.

LEMMA 3.1. *There exist a constant $K_2>0$ and a $\delta_1>0$, both of which depend only on $\mathscr{K}$, such that if (τ,ξ) and $(\bar\tau,\bar\xi)$ are in $\mathscr{K}\cap\mathscr{R}$ and $|(\tau,\xi)-(\bar\tau,\bar\xi)|<\delta_1$, then*

$$|t_1-\bar t_1|\leqq K_2(|\tau-\bar\tau|+|\xi-\bar\xi|). \tag{3.4}$$

We first note that if both t_1 and $\bar t_1$ equal T, then (3.4) is trivially true. We therefore suppose henceforth that at least one of the inequalities $t_1<T$ or $\bar t_1<T$ holds.

By (2.2) of Lemma 2.1 there exists an $R>0$ that depends on $\mathscr{K}$ such that the terminal point $(t_1,\phi(t_1))$ of any trajectory with initial point (τ,ξ) in $\mathscr{K}$ will lie in $([0,T]\times B_R)\cap\partial\mathscr{T}$, where B_R denotes the closed ball of radius R in $\mathbb{R}^n$. Let $\mathscr{B}_R=([0,T]\times B_R)\cap\hat\partial\mathscr{F}$. Then $\mathscr{B}_R$ is a compact subset of $\hat\partial\mathscr{F}$. Also, any trajectory ϕ with initial point in $\mathscr{K}$ whose terminal time t_1 is less than T satisfies $(t_1,\phi(t_1))\in\mathscr{B}_R$.

Let $\varepsilon_1=\varepsilon_1(\mathscr{B}_R)$ and $c=c(\mathscr{B}_R)$ be the constants in Assumption II associated with this $\mathscr{B}_R$.

By (2.4) of Lemma 2.1 there exists a $\delta_1>0$ such that if $|(\tau,\xi)-(\bar\tau,\bar\xi)|<\delta_1$, then for all $\max(\tau,\bar\tau)\leqq t\leqq T$,

$$|\phi(t,\tau,\xi)-\bar\phi(t,\bar\tau,\bar\xi)|<\varepsilon_1/2.$$

We now suppose that $|(\tau,\xi)-(\bar\tau,\bar\xi)|<\delta_1$ and that $t_1<\bar t_1$. Then $t_1<T$ and $(t_1,\phi(t_1))\in\mathscr{B}_R$. Since $|(t_1,\phi(t_1))-(t_1,\bar\phi(t_1))|=|\phi(t_1)-\bar\phi(t_1)|<\varepsilon_1/2$, it follows that $(t_1,\bar\phi(t_1))\in\mathscr{N}_{\varepsilon_1/2}(\mathscr{B}_R)$.

Let $\bar\theta(t)=\rho(t,\bar\phi(t))$. Then $\bar\theta$ is absolutely continuous and

$$\frac{d\bar\theta}{dt}=\rho_t(t,\bar\phi(t))+\langle\rho_x(t,\bar\phi(t)),f(t,\bar\phi(t),\bar u(t))\rangle \quad \text{a.e.} \tag{3.5}$$

From (3.3) we see that as long as $(t,\bar\phi(t))$ stays in $\mathscr{N}_{\varepsilon_1}(\mathscr{B}_R)$, the right-hand side of (3.5) does not exceed $-c$. Since $(t_1,\bar\phi(t_1))\in\mathscr{N}_{\varepsilon_1/2}(\mathscr{B}_R)$, it follows from the continuity of $\bar\phi$ that there exists a maximal interval $[t_1,t_1+\alpha)$ with $t_1+\alpha\leqq\bar t_1$ such that if t is in $[t_1,t_1+\alpha)$, then $(t,\bar\phi(t))\in\mathscr{N}_{\varepsilon_1}(\mathscr{B}_R)$ and $(t,\bar\phi(t))\notin\mathscr{T}$. For all t in this interval we have

$$\bar\theta(t)-\bar\theta(t_1)\leqq -c(t-t_1). \tag{3.6}$$

From (3.6) we first see that $t_1+\alpha<\bar t_1$ is not possible. For then, since $\bar\theta(t_1)<\varepsilon_1/2$, and $\bar\theta(t_1+\alpha)=\varepsilon_1$, we would have $c\alpha\leqq-\varepsilon_1/2$, which is impossible. Hence (3.6) holds for $t_1\leqq t\leqq\bar t_1$. If $\bar t_1<T$, then $\bar\theta(\bar t)=0$ and we get from (3.6) that

$$(\bar t_1-t_1)\leqq c^{-1}\bar\theta(t_1)\leqq c^{-1}|(t_1,\phi(t_1))-(t_1,\bar\phi(t_1))|=c^{-1}|\phi(t_1)-\bar\phi(t_1)|.$$

It now follows from (2.4) that

$$(\bar t_1-t_1)\leqq K_2(|\tau-\bar\tau|+|\xi-\bar\xi|) \tag{3.7}$$

where $K_2=K_1/c$. If $\bar t_1=T$, then $\bar\theta(T)\geqq 0$, in which case (3.7) still holds.

If $\bar{t}_1 < t_1$, we reverse the roles of t_1 and $\bar{t}$, and obtain that (3.7) holds with left-hand side equal to $t_1 - \bar{t}_1$. From these two inequalities the lemma follows.

LEMMA 3.2. *With the notation as above, there exists a constant $K_3 > 0$ that depends only on $\mathscr{X}$, such that for (τ, ξ) and $(\bar{\tau}, \bar{\xi})$ in $\mathscr{X} \cap \mathscr{R}$,*

$$|\bar{\phi}(\bar{t}_1) - \phi(t_1)| \leqq K_3(|\tau - \bar{\tau}| + |\xi - \bar{\xi}|)$$

whenever $|(\tau, \xi) - (\bar{\tau}, \bar{\xi})| \leqq \delta_1$.

Proof. $|\bar{\phi}(\bar{t}_1) - \phi(t_1)| \leqq |\bar{\phi}(\bar{t}_1) - \phi(\bar{t}_1)| + |\phi(\bar{t}_1) - \phi(t_1)|$. By (2.3) and (2.4) of Lemma 2.1, the right-hand side of the preceding inequality does not exceed $K_1(|\tau - \bar{\tau}| + |\xi - \bar{\xi}|) + (1 + R)|\bar{t}_1 - t_1|$. Lemma 3.2 now follows from this and from Lemma 3.1.

From (5) of Assumption I and the two preceding lemmas we obtain the following:

$$\begin{aligned} g(\bar{t}_1, \bar{\phi}(\bar{t}_1)) - g(t_1, \phi(t_1)) &\leqq K_R(|t_1 - \bar{t}_1| + |\phi(t_1) - \bar{\phi}(\bar{t}_1)|) \\ &\leqq K_3(|\tau - \bar{\tau}| + |\xi + \bar{\xi}|). \end{aligned}$$

Recall that $\bar{u}$ depends on u. Hence

$$W(\bar{\tau}, \bar{\xi}) \leqq W(\tau, \xi) + K_3(|\tau - \bar{\tau}| + |\xi - \bar{\xi}|).$$

Reversing the roles of (τ, ξ) and $(\bar{\tau}, \bar{\xi})$ gives (3.2) with $K = K_3$ when both (τ, ξ) and $(\bar{\tau}, \bar{\xi})$ are in $\mathscr{X} \cap \mathscr{R}$.

We now suppose that $(\tau, \xi) \in \mathscr{X} \cap \mathscr{R}$ and that $(\bar{\tau}, \bar{\xi}) \in \mathscr{X} \cap \partial\mathscr{T}$. To emphasize that $(\bar{\tau}, \bar{\xi}) \in \partial\mathscr{T}$ we shall write this point as $(\bar{t}_1, \bar{x}_1)$.

We first consider the case in which $\bar{t}_1 = T$.

As above, let (t_1, x_1) be the terminal point of $\phi(\cdot) = \phi(\ , \tau, \xi, u)$, where u is a control in $[\tau, T]$. Then

$$0 \leqq \bar{t}_1 - t_1 = T - t_1 < T - \tau = \bar{t}_1 - \tau = \bar{\tau} - \tau. \tag{3.8}$$

Also,

$$\begin{aligned} |x_1 - \bar{x}_1| &= |\xi + \int_\tau^{t_1} f(s, \phi(s), u(s))\, ds - \bar{x}_1| \\ &\leqq |\xi - \bar{x}_1| + c'|\bar{t}_1 - \tau| \end{aligned} \tag{3.9}$$

for some constant $c' > 0$.

We now consider the case $\bar{t}_1 \neq T$. Then $(\bar{t}_1, \bar{x}_1) \in \hat{\partial}\mathscr{F}$. Since $(\bar{t}_1, \bar{x}_1) = (\bar{\tau}, \bar{\xi})$ belongs to $\mathscr{X}$ we also have that $(\bar{t}_1, \bar{x}_1) \in \mathscr{B}_R$. Let ε_1 and c be as before. We take (τ, ξ) to be such that $(\tau, \xi) \in \mathscr{N}_{\varepsilon_1/2}(\mathscr{B}_R)$ and $|(\tau, \xi) - (\bar{\tau}, \bar{\xi})| < \delta_1$.

Let $\theta(t) = \rho(t, \phi(t))$ for $\tau \leqq t \leqq t_1$ and note that $t_1 \leqq T$. Then arguing as before, we get that $\theta(t) - \theta(\tau) \leqq -c(t - \tau)$ for all $\tau \leqq t \leqq t_1$. If $t_1 < T$, then $\theta(t_1) = 0$ and we get

$$0 < t_1 - \tau \leqq c^{-1}\theta(\tau) \leqq c^{-1}(|\tau - \bar{t}_1| + |\xi - \bar{x}_1|). \tag{3.10}$$

If $t_1 = T$, we get

$$0 < t_1 - \tau \leqq c^{-1}(\theta(\tau) - \theta(T)) \leqq c^{-1}\theta(\tau),$$

so that (3.10) still holds. Hence we always have

$$|t_1 - \bar{t}_1| \leqq |t_1 - \tau| + |\tau - \bar{t}_1| \leqq A(|\tau - \bar{t}_1| + |\xi - \bar{x}_1|) \tag{3.11}$$

for some constant $A > 0$. Comparing (3.11) with (3.8), we see that (3.11) holds whether or not $\bar{t}_1 = T$.

Also, by (3) of Assumption I and (3.10),

$$\begin{aligned} |x_1 - \bar{x}_1| &\leqq |x_1 - \xi| + |\xi - \bar{x}_1| \\ &\leqq \int_\tau^{t_1} |f(s, \phi(s), u(s))\, ds\,| + |\xi - \bar{x}_1| \\ &\leqq B|t_1 - \tau| + |\xi - \bar{x}_1| \\ &\leqq C(|\tau - \bar{t}_1| + |\xi - \bar{x}_1|). \end{aligned} \tag{3.12}$$

Comparing (3.12) with (3.9) we see that (3.12) also holds whether or not $\bar{t}_1 = T$.

It now follows from (3.11) and (3.12) and (5) of Assumption I, that if $|(\tau, \xi) - (\bar{t}_1, \bar{x})| < \delta$, where $\delta \equiv \min(\varepsilon_1/2, \delta_1)$, then there exists a $K_4 > 0$ such that

$$\begin{aligned} g(t_1, x_1) - W(\bar{t}_1, \bar{x}_1) &= g(t_1, x_1) - g(\bar{t}_1, \bar{x}_1) \\ &\leqq K'_R(|t_1 - \bar{t}_1| + |\bar{x}_1 - x_1|) \\ &\leqq K_4(|\tau - \bar{t}_1| + |\xi - \bar{x}|). \end{aligned}$$

From this inequality (3.2) follows with $K = K_4$.

If both (τ, ξ) and $(\bar{\tau}, \bar{\xi})$ belong to $\partial\mathcal{T}$ then (3.2) follows with $K = K'_R$ from (3.1) and (5) of Assumption I.

Hence the theorem holds with $\delta = \min(\varepsilon_1/2, \delta_1)$ and $K = \max(K_3, K_4, K'_R)$.

4. A necessary condition. The principal result of this section is the following theorem.

THEOREM 4.1. *Let Assumptions* I *and* II *hold and for each* $0 \leqq \tau < T$ *let there exist a control* u *on* $[\tau, T]$. *For each* τ *in* $[0, T]$ *and each* z *in* $Z(t)$ *let there exist a* $\delta_0 > 0$ *and a control* u *in* $[\tau, \tau + \delta_0)$ *such that* $\lim_{t \to \tau+0} u(t) = z$. *Then for each* (τ, ξ) *in* $\mathcal{R}$,

$$\min_{h \in Q(\tau, \xi)} D^- W(\tau, \xi; 1, h) = 0. \tag{4.1}$$

Before we prove Theorem 4.1 we discuss some of its implications.

We have already pointed out that Assumptions I and II imply that if for each τ the set of controls on $[\tau, T]$ is not empty, then Problem I with initial point (τ, ξ) in $\mathcal{R}$ has a solution. Thus W is well defined on $\mathcal{R}$. In Theorem 3.1 we have shown that W is locally Lipschitz in $\mathcal{R} \cup \mathcal{T}$. Thus W is differentiable almost everywhere on $\mathcal{R}$. At points of differentiability (4.1) becomes

$$W_t(\tau, \xi) + \min_{z \in Z(t)} \langle W_t(\tau, \xi), f(t, x, z) \rangle = 0 \tag{4.2}$$

where $W_x = (\partial W/\partial x^1, \cdots, \partial W/\partial x^n)$.

Let (τ, ξ) be a point in $\mathcal{R}$ and let $\psi(\ ) = \psi(\ , \tau, \xi)$ denote an optimal trajectory for the problem with initial point (τ, ξ). Since W is Lipschitz continuous and ψ is absolutely continuous, the function ω defined on $[\tau, t_1]$ by the formula $\omega(t) = W(t, \psi(t))$ is absolutely continuous. Thus almost all points of $[\tau, t_1]$ are simultaneously

Lebesgue points of ψ and points of differentiability of ω. At such a point we have

$$\begin{aligned}\frac{d\omega}{dt} &= \lim_{\delta\to 0}[\,W(t+\delta,\psi(t+\delta))-W(t,\psi(t))]\delta^{-1}\\ &= \lim_{\delta\to 0}\left[\,W\left(t+\delta,\psi(t)+\int_t^{t+\delta}\psi'(s)\,ds\right)-W(t,\psi(t))\right]\delta^{-1}\\ &= \lim_{\delta\to 0}[\,W(t+\delta,\psi(t)+\delta\psi'(t)+o(\delta))-W(t,\psi(t))]\delta^{-1}\\ &= DW(t,\psi(t);1,\psi'(t))\end{aligned}$$

where in passing to the last line we have used the Lipschitz continuity of W.

On the other hand, the Principle of Optimality, discussed below, gives the relation

$$W(t+\delta,\psi(t+\delta))-W(t,\psi(t))=0.$$

Hence $DW(t,\psi(t);1,\psi'(t))=0$. Combining this relation with (4.1) gives the following result that can be considered as a form of the Pontryagin maximum principle.

COROLLARY 4.1. *At almost all points of* $[\tau,t_1]$,

$$0=DW(t,\psi(t);1,\psi'(t))\leqq DW^-(t,\psi(t);1,h)$$

for all $h\in Q(t,\psi(t))$.

If for each t in $[0,T]$ the set $Z(t)$ is compact, or if we strengthen (3) of Assumption I to require that β be finite-valued for all t and that (3) holds for all t, then

$$H(t,x,p)=\min_{h\in Q(t,x)}\langle p,h\rangle=\min_{z\in Z(t)}\langle p,f(t,x,z)\rangle$$

is defined and finite on $[0,T]\times\mathbb{R}^n\times\mathbb{R}^n$. In this case, an elementary argument due to Lions and Souganidis [10] and based on one of the equivalent definitions of viscosity solution shows that (4.1) implies that W is a viscosity solution of

$$\begin{aligned}&u_t+H(t,x,u_x)=0, && (t,x)\in\mathcal{R},\\ &u(t,x)=g(t,x), && (t,x)\in\partial\mathcal{T}.\end{aligned}\tag{4.3}$$

The next theorem is the Principle of Optimality for Problem I. Note that we do not require Assumption II to hold for the terminal set.

THEOREM 4.2. *Let* $\mathcal{R}$ *be a connected set in* $[0,T)\times\mathbb{R}^n$ *and let* $\mathcal{T}'$ *be a set in* $[0,T]\times\mathbb{R}^n$ *such that* $\partial\mathcal{T}'\subseteq\partial\mathcal{R}$. *For each* (τ,ξ) *in* $\mathcal{R}$ *let Problem* I *with terminal set* $\mathcal{T}'$ *have a solution* $(u^*(\cdot),\phi^*(\cdot))$ *where* $u^*(\cdot)=u^*(\ ,\tau,\xi)$ *and* $\phi^*(\cdot)=\phi^*(\ ,\tau,\xi)$. *Then if* t_f^* *is the terminal point of* ϕ^*,

$$W(t,\phi^*(t))=W(\tau,\xi),\qquad \tau\leqq t\leqq t_f^*.\tag{4.4}$$

If (u,ϕ) *is any other control trajectory pair with initial point* (τ,ξ) *and terminal point* t_f, *then*

$$W(t,\phi(t))\geqq W(\tau,\xi),\qquad \tau\leqq t\leqq t_f.\tag{4.5}$$

The proof is well known and will be omitted.

We shall also need the following result, whose proof we omit.

LEMMA 4.1. *Let* K *be a set in* $\mathbb{R}^n$ *of positive finite Lebesgue measure and let* $\lambda: K\to\mathbb{R}^n$ *be Lebesgue integrable on* K. *Let* $\mathcal{H}$ *denote the set of regular probability measures* μ *such that* $\int_K f\,d\mu$ *exists. Then the sets* $\mathcal{P}=\operatorname{clco}\{f(x): x\in K\}$ *and* $\mathcal{P}_\mu=\operatorname{cl}\{\int_K f\,d\mu: \mu\in\mathcal{H}\}$ *are equal.*

We now take up the proof of Theorem 4.1. We first show that

$$\inf_{h\in Q(\tau,\xi)} D^-W(\tau,\xi;1,h)\geqq 0.\tag{4.6}$$

Let $h \in Q(\tau, \xi)$. Then there exists a z in $Z(\tau)$ such that $h = f(\tau, \xi, z)$. Also, there exists a control u defined on $[\tau, \tau+\delta_0)$ with $u(\tau) = z$ that is continuous from the right at $t = \tau$. This control can be extended to $[\tau, T]$, and we denote the extended control also by u. Let ϕ be the trajectory corresponding to u and having initial point (τ, ξ). Then for $\delta > 0$

$$\begin{aligned} \phi(\tau+\delta) &= \xi + \int_{\tau}^{\tau+\delta} f(s, \phi(s), u(s))\, ds \\ &= \xi + \int_{\tau}^{\tau+\delta} [f(\tau, \xi, z) + o(1)]\, ds \\ &= \xi + \delta f(\tau, \xi, z) + o(\delta), \end{aligned} \tag{4.7}$$

where o is as $\delta \to 0$.

By the Principle of Optimality,

$$[W(\tau+\delta, \phi(\tau+\delta)) - W(\tau, \xi)]\delta^{-1} \geqq 0.$$

If we substitute the rightmost expression of (4.7) into this inequality and use the Lipschitz continuity of W, we get that

$$[W(\tau+\delta, \xi+\delta h) - W(\tau, \xi)]\delta^{-1} + o(1) \geqq 0.$$

From this, (4.6) follows.

We next show that there exists an $h^* \in Q(\tau, \xi)$ such that $D^- W(\tau, \xi; 1, h^*) \leqq 0$. This in conjunction with (4.6) will establish the theorem.

Let ψ denote the optimal trajectory for the problem with initial point (τ, ξ). Then

$$\psi(\tau+\delta) = \xi + \int_{\tau}^{\tau+\delta} \psi'(s)\, ds \tag{4.8}$$

where $\psi'(s) \in Q(s, \psi(s))$ for almost all $\tau \leqq s \leqq T$. Hence, since ψ is continuous, given an $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that if $\tau \leqq s \leqq \tau+\delta$ then

$$\psi'(s) \in Q(\mathcal{N}_\varepsilon(\tau, \xi)) \quad \text{a.e.} \tag{4.9}$$

Let K_ε denote the set of points in $[\tau, \tau+\delta]$ at which the inclusion (4.9) holds. Then the measure of K_ε equals δ. Thus

$$\int_{\tau}^{\tau+\delta} \psi'(s)\, ds = \delta \int_{\tau}^{\tau+\delta} \psi'(s)\left(\frac{ds}{\delta}\right) = \delta \int_{K_\varepsilon} \psi'(s)\left(\frac{ds}{\delta}\right).$$

From (4.9) we get that

$$\operatorname{clco}\{\psi'(s): s \in K_\varepsilon\} \subseteq \operatorname{clco}\{Q(\mathcal{N}_\varepsilon(\tau, \xi))\}.$$

From this relationship and from Lemma 4.1 we get that

$$\int_{K_\varepsilon} \psi'(s)\left(\frac{ds}{\delta}\right) \in \operatorname{clco}\{Q(\mathcal{N}_\varepsilon(\tau, \xi))\}.$$

Let

$$h_\delta \equiv \int_{\tau}^{\tau+\delta} \psi'(s)\left(\frac{ds}{\delta}\right).$$

We have shown that for every $\varepsilon > 0$, there exists a $\delta > 0$ and a point h_δ such that

$$\psi(\tau+\delta) = \xi + \delta h_\delta, \qquad h_\delta \in \operatorname{clco}(\mathcal{N}_\varepsilon(\tau, \xi)). \tag{4.10}$$

Since

$$|h_\delta| \leqq \delta^{-1} \int_{\tau}^{\tau+\delta} |\psi'(s)|\, ds \leqq \delta^{-1} \int_{\tau}^{\tau+\delta} |1 + \psi(s)|\, \|\beta\|_\infty\, ds$$

where $\|\beta\|_\infty$ is the $L_\infty[0, T]$ norm of β, it follows from (2.2) of Lemma 2.1 that there exists a constant A such that $|h_\delta| \leqq A$ for all $\delta < \delta(\varepsilon)$. Hence there exist sequences $\varepsilon_k \to 0$ and $\delta_k = \delta_k(\varepsilon_k) \to 0$ such that $h_k \in \operatorname{clco}\{Q(\mathcal{N}_{\varepsilon_k}(\tau, \xi))\}$ and such that h_k converges to some element h^* in $\mathbb{R}^n$.

Let $\varepsilon > 0$ be fixed, but arbitrary. Then there exists an integer k' such that if $k > k'$, then $Q(\mathcal{N}_{\varepsilon_k}(\tau, \xi)) \subseteq Q(\mathcal{N}_\varepsilon(\tau, \xi))$. Hence $\operatorname{clco}\{Q(\mathcal{N}_{\varepsilon_k}(\tau, \xi))\} \subseteq \operatorname{clco}\{Q(\mathcal{N}_\varepsilon(\tau, \xi))\}$, and so $h_k \in \operatorname{clco}\{Q(\mathcal{N}_\varepsilon(\tau, \xi))\}$. Hence, $h^* \in \operatorname{clco}\{Q(\mathcal{N}_\varepsilon(\tau, \xi))\}$ for arbitrary $\varepsilon > 0$. It now follows from property (Q) that $h^* \in Q(\tau, \xi)$.

We return to our sequences $\{\delta_k\}$ and $\{h_k\}$. From the definition of h^* and from (4.10) we get that

$$\psi(\tau + \delta_k) = \xi + \delta_k h^* + o(\delta_k). \tag{4.11}$$

From the Principle of Optimality we have that

$$[W(\tau + \delta_k, \psi(\tau + \delta_k)) - W(\tau, \xi)]\delta_k^{-1} = 0.$$

Substituting (4.11) into this relation and using the Lipschitz continuity of W, we get that

$$\lim_{k \to \infty} [W(\tau + \delta, \xi + \delta_k h^*) - W(\tau, \xi)]\delta_k^{-1} = 0.$$

Recall that $\delta_k \to 0$ as $k \to \infty$. Hence

$$\liminf_{\delta \to 0+} [W(\tau + \delta, \xi + \delta h^*) - W(\tau, \xi)]\delta^{-1} \leqq 0,$$

so that $D^- W(\tau, \xi; 1, h^*) \leqq 0$. This proves the theorem. Note that our argument shows that $D^- W(\tau, \xi; 1, h^*) = 0$, so that we are justified in writing min in (4.1).

5. Optimal synthesis. Let V be a real-valued function that is continuous on $\mathcal{R} \cup \partial\mathcal{T}$, that is locally Lipschitz on $\mathcal{R}$, and that satisfies

$$\begin{aligned} &\min_{h \in Q(t, x)} D^- V(t, x; 1, h) = 0, && (t, x) \in \mathcal{R}, \\ &V(t_1, x_1) = g(t_1, x_1), && (t_1, x_1) \in \partial\mathcal{T}. \end{aligned} \tag{5.1}$$

We have shown in Theorems 3.1 and 4.1 that if the assumptions of Theorem 4.1 hold, then the value function W is one such function. In the case of compact constraints, or in the case that (3) of Assumption I is strengthened to require that β be finite-valued for all t and that (3) holds for all t, any viscosity solution of (4.3) that is continuous on $\mathcal{R} \cup \mathcal{T}$ and locally Lipschitz continuous on $\mathcal{R}$ can be shown to satisfy (5.1). The proof of this assertion makes use of the definition of viscosity solution and arguments similar to, but simpler than, those of Theorem 4.1. No uniqueness theorems for viscosity solutions are used.

For each (t, x) in $\mathcal{R}$ let

$$F(t, x) = \{h: h \in Q(t, x), D^- V(t, x; 1, h) = 0\}. \tag{5.2}$$

The set-valued function F furnishes a generalized optimal synthesis or feedback in $\mathcal{R}$ in the following sense.

LEMMA 5.1. *For each (τ, ξ) in $\mathcal{R}$, any solution of the differential inclusion*

$$x' = F(t, x), \qquad x(\tau) = \xi \tag{5.3}$$

that is defined on an interval $[\tau, \tau + \alpha)$ such that ψ hits $\mathcal{T}$ at a time $t_1 < \alpha$ is an optimal trajectory for Problem I *with initial point (τ, ξ).*

Let ψ be a solution of (5.3) with the hypothesized properties. Since $F(t, x) \subseteq Q(t, x)$, the solution ψ also satisfies $\psi'(t) \in Q(t, \psi(t))$ almost everywhere. By Filippov's Lemma there exists a control v such that $\psi'(t) = f(t, \psi(t), v(t))$ almost everywhere. Thus v and ψ are a control and trajectory pair.

To see that (ψ, v) is optimal, let (ϕ, u) be any control trajectory pair for the problem with initial point (τ, ξ). Then the function Φ on $[\tau, t_1]$, defined by $\Phi(t) = V(t, \phi(t))$, is absolutely continuous on $[\tau, t_1]$. Hence

$$
\begin{aligned}
V(t_1, \phi(t_1)) - V(\tau, \xi) &= \int_\tau^{t_1} \left[\frac{d}{ds} W(s, \phi(s)) \right] ds \\
&= \int_\tau^{t_1} DV(s, \phi(s); 1, \phi'(s))\, ds \geqq 0.
\end{aligned} \tag{5.4}
$$

The second equality follows by the same argument that was used to calculate dw/dt in the proof of Corollary 4.1 and the last inequality follows from the first condition in (5.1). From (5.4) and the second condition in (5.1) we get that

$$
g(t_1, \phi(t_1)) \geqq V(\tau, \xi). \tag{5.5}
$$

If we take $(\phi, u) = (\psi, v)$ in the preceding argument, then the last inequality in (5.4) will be replaced by an equality. Hence we get $g(\bar{t}_1, \psi(\bar{t}_1)) = V(\tau, \xi)$, where $\bar{t}_1$ is the terminal time of ψ. If we combine this relation with (5.5) we get that for all control trajectory pairs (ϕ, u) for the problem with initial point (τ, ξ), $g(t_1, \phi(t_1)) \geqq g(\bar{t}_1, \psi(\bar{t}_1))$.

COROLLARY 5.1. *If for each (τ, ξ) in $\mathcal{R}$ equation (5.3) has a solution $\psi(\ , \tau, \xi)$ on an interval sufficiently large for ψ to intersect $\mathcal{T}$, then $V = W$.*

Remark 5.1. We have already noted that in the case of compact constraints or a suitably strengthened (3) of Assumption I, a viscosity solution of (4.3) that is locally Lipschitz continuous on $\mathcal{R}$ satisfies (5.1). Thus, such a solution of (4.3) will give a synthesis of optimal control in the sense indicated.

We now define a sequence of functions that will converge to a solution of (5.3) when F satisfies certain conditions. This will give a constructive procedure for finding solutions that, in principle, can be implemented to obtain approximations to solutions.

For each $n = 1, 2, 3, \cdots$ we define a Euler polygon approximation E_n of order n for (5.3). To emphasize the dependence of E_n on the initial point (τ, ξ) we write $E_n(\ , \tau, \xi)$. Let π_n be a partition of $[\tau, T]$ with partition points $\tau = \tau_0 < \tau_1 < \cdots < \tau_{2n} = T$ with the following properties. (i) For $i > 0$, $|\beta(\tau_i)| \leqq \|\beta\|_\infty$, where β is the function in (3) of Assumption I. (ii) For $i > 0$, (3) of Assumption I holds with $t = \tau_i$. (iii) If $\delta_n \equiv \max\{0 \leqq i \leqq 2n-1\colon \tau_{i+1} - \tau_i\}$, then $\delta_n \leqq (T - \tau)/n$.

Let h_0 be an element of $F(\tau_0, \xi_0)$, which is fixed for a particular sequence of Euler polygons. For $\tau_0 \leqq t \leqq \tau_1$ define

$$
E_n(t) = \xi_0 + h_0(t - \tau_0). \tag{5.6}
$$

Now suppose that $E_n(t)$ has been defined for $\tau_0 \leqq t \leqq \tau_i$, where $1 \leqq i < 2n-1$. Let $\xi_i = E_n(\tau_i)$, and let h_i be any element of $F(\tau_i, \xi_i)$. Then for $\tau_i \leqq t \leqq \tau_{i+1}$ define

$$
E_n(t) = \xi_i + h_i(t - \tau_i). \tag{5.7}
$$

Thus, E_n is defined for all $\tau_0 \leqq t \leqq T$. Note that the triple (τ_0, ξ_0, h_0) is independent of n, but that for $i > 0$, the triple (τ_i, ξ_i, h_i) depends on n.

LEMMA 5.2. *Given $(\tau, \xi) \in \mathcal{R}$ and a choice $h_0 \in F(\tau, \xi)$, then there exist constants $R > 0$ and $M > 0$ such that for all all n and all $\tau \leqq t \leqq T$, we have $|E_n(t)| \leqq R$ and $|E_n'(t)| \leqq M$. At the partition points τ_i, $E_n'(\tau_i)$ is interpreted as either of the one-sided derivatives.*

To establish $|E_n(t)| \leqq R$ it suffices to show that $|\xi_i| \leqq R$ for all n and all $0 \leqq i \leqq 2n$, where we take $\xi_{2n} = E_n(T)$. Let $R_1 = \max(|h_0|, \|\beta\|_\infty)$. It then follows from (5.6) and (3) of Assumption I that $1+|\xi_1| \leqq (1+|\xi_0|)(1+R_1\delta_n)$. By a simple induction argument, (5.7) and (3) of Assumption I, we get that for all n and all $0 \leqq i \leqq 2n-1$,

$$1+|\xi_{i+1}| \leqq (1+|\xi_0|)(1+TR_1n^{-1})^{2n}.$$

From this it follows that there exists an $R>0$ such that $|\xi_i| \leqq R$ for all n and all $0 \leqq i \leqq 2n-1$.

For $\tau \leqq t \leqq T$, $E_n'(t) = h_i$ for appropriate $0 \leqq i \leqq 2n$. By construction and (3) of Assumption I, for $i>0$ we have $|h_i| \leqq \|\beta\|_\infty(1+|\xi_i|)$. Since h_0 is independent of n and $|\xi_i| \leqq R$ for all n and $0 \leqq i \leqq 2n$, the relation $|E_n'(t)| \leqq M$ follows.

The functions $\{E_n\}$ are uniformly bounded and equicontinuous, so there exists a subsequence that we again label as $\{E_n\}$, that converges uniformly on $[\tau, T]$ to a continuous function ψ. Since the integrals $\int_\tau^t E_n'(s)\,ds$, $\tau \leqq t \leqq T$ are uniformly bounded and uniformly absolutely continuous, the function ψ is absolutely continuous and the derivatives E_n' converge weakly in $L_1[\tau, T]$ to ψ'. If we could show that $\psi'(t) \in F(t, \psi(t))$ almost everywhere on $[\tau, t_1]$, where t_1 is the terminal time, then by Lemma 5.1, ψ would be optimal. Moreover, the convergent subsequence $\{E_n\}$ would approximate ψ uniformly.

We now take up conditions that are sufficient for ψ to be a solution of (5.3).

DEFINITION 5.1. A subset $\mathscr{E}$ of $\mathscr{R}$ is said to be an exceptional set for the synthesis if for each (τ, ξ) in $\mathscr{R}$ there exists a limit point $\psi(\ , \tau, \xi)$ in $C^n[\tau, T]$ of some sequence $\{E_n(\ , \tau, \xi)\}$ of Euler polygons such that the set

$$\mathscr{J}(\psi) = \{\tau \leqq t \leqq t_1 : (t, \psi(t)) \in \mathscr{E}\} \tag{5.8}$$

has Lebesgue measure zero. Here t_1 is the terminal point of ψ.

Remark 5.2. Any set $\mathscr{E}_0$ whose projection on $[0, T]$ has measure zero can be an exceptional set. Exceptional sets also arise as "dispersal surfaces," in the terminology of Isaacs [9] for control problems. Example 5.1 below illustrates a simple case of this.

THEOREM 5.1. *Let the mapping F satisfy property (Q) at each point (t, x) of $\mathscr{R}$, with the possible exception of points in an exceptional set $\mathscr{E}$. Then for each (τ, ξ) in $\mathscr{R}$ any uniform limit $\psi(\ , \tau, \xi)$ of a sequence $E_n(\tau, \xi)$ of Euler polygons that satisfies* meas $\mathscr{J}(\psi) = 0$ *satisfies* (5.3) *almost everywhere on $[\tau, T]$ and is optimal.*

Proof. In view of the discussion preceding Definition 5.1, we need only prove that ψ satisfies (5.3).

We have already shown that $E_n' \to \psi'$ weakly in $L_1[\tau, T]$. Hence by Mazur's Theorem there exists a sequence of functions ψ_{n_j} defined by the formulas

$$\psi_{n_j} = \sum_{i=1}^{k(n_j)} \alpha_{n_j,i} E_{n_j+i}, \quad \alpha_{n_j,i} \geqq 0, \qquad \sum_{i=1}^{k(n_j)} \alpha_{n_j,i} = 1$$

where $n_{j+1} > n_j + k(n_j)$, and such that $\psi_{n_j}' \to \psi'$ in $L_1[\tau, T]$. An elementary argument shows that the uniform convergence of E_n to ψ implies that $\psi_{n_j} \to \psi$ uniformly on $[\tau, T]$. Since $\psi_{n_j}' \to \psi'$ in $L_1[\tau, T]$, there exists a subsequence that we again label as ψ_{n_j} such that $\psi_{n_j}' \to \psi'$ almost everywhere.

Let $t_0 > \tau$ be a point of $[\tau, T]$ such that $\psi_{n_j}'(t_0) \to \psi'(t_0)$ and $t_0 \notin \mathscr{J}(\psi)$. The set of such points has full measure. We conclude the proof by showing at such a point $\psi'(t_0) \in F(t_0, \psi(t_0))$.

Let $\delta > 0$ be given. Since $|E_n'(t)| \leqq M$ for all t in $[\tau, T]$ and all n, there exists a positive η such that $\eta < \delta/3$ and if $|t'-t''| < \eta$, then $|E_n(t') - E_n(t'')| < \delta/3$ for all n. Also, there exists a positive integer N_1 such that for $n > N_1$, $|E_n(t) - \psi(t)| < \delta/3$ for all t in $[\tau, T]$.

There exists a positive integer N_2 such that for all $n > N_2$, the partition π_n has partition points in $(t_0 - \eta, t_0]$. Let $\nu = \nu(n)$ be the index such that τ_ν is a partition point of π_n in $(t_0 - \eta, t_0]$ and satisfies $\tau_\nu \leqq t_0 < \tau_{\nu+1}$.

Let $N = \max(N_1, N_2)$. Then for $n > N$, (recall that $\nu = \nu(n)$)

$$|(\tau_\nu, E_n(\tau_\nu)) - (t_0, \psi(t_0))| \leqq |\tau_\nu - t_0| + |E_n(\tau_\nu) - \psi(t_0)|$$

$$\leqq |\tau_\nu - t_0| + |E_n(\tau_\nu) - E_n(t_0)| + |E_n(t_0) - \psi(t_0)| < \delta.$$

In other words, we have $(\tau_\nu, E_n(\tau_\nu)) \in \mathcal{N}_\delta(t_0, \psi(t_0))$. By our definition of E_n, we have that $E_n'(\tau_\nu + 0) \in F(\tau_\nu, E_n(\tau_\nu))$. Hence, $E_n'(\tau_\nu + 0) \in F(\mathcal{N}_\delta(t_0, \psi(t_0)))$. But by our choice of ν and the definition of E_n, we have that $E_n'(\tau_\nu + 0) = E_n'(t_0)$. Hence $E_n'(t_0) \in F(\mathcal{N}_\delta(t_0, \psi(t_0)))$, for $n > N$. This in turn implies that there exists a positive integer $J = J(\delta)$ such that for $j > J$,

$$\psi_{n_j}'(t_0) \in \operatorname{co}\{F(\mathcal{N}_\delta(t_0, \psi(t_0)))\}.$$

Since $\psi_{n_j}'(t_0) \to \psi'(t_0)$, the preceding statement implies that $\psi'(t_0) \in \operatorname{clco}\{F(\mathcal{N}_\delta(t_0, \psi(t_0)))\}$. Since $\delta > 0$ was arbitrary

$$\psi'(t_0) \in \bigcap_{\delta > 0} \operatorname{clco}\{F(\mathcal{N}_\delta(t_0, \psi(t_0)))\}.$$

Finally since property (Q) holds at $(t_0, \psi(t_0))$, we get that $\psi'(t_0) \in F(t_0, \psi(t_0))$, and the theorem is proved.

We illustrate Theorem 5.1 with a simple example.

Example 5.1. Let $n = 1$, let $T = 1$, let the terminal set be $[1, \infty) \times \mathbb{R}$, and let $g(t, x) = -x^2$. We wish to minimize $g(1, \phi(1)) = -(\phi(1))^2$, over all ϕ that satisfy $x' = u$, $x(\tau) = \xi$, and $-1 \leqq u \leqq 1$.

It is clear that for (τ, ξ) with $\xi > 0$, the optimal control is $u(t) = 1$ for $\tau \leqq t \leqq 1$ and the optimal trajectory is the straight line of slope one through (τ, ξ). Similarly for (τ, ξ) with $\xi < 0$, the optimal control is $u(t) = -1$ and the optimal trajectory is the straight line of slope negative one. For (τ, ξ) with $\xi = 0$, both $u(t) = +1$ and $u(t) = -1$ are optimal controls and both the lines with slope plus one and minus one are optimal. The value function W is given by $W(\tau, \xi) = -[\xi + (1 - \tau)]^2$ if $\xi \geqq 0$ and by $W(\tau, \xi) = -[\xi - (1 - \tau)]^2$ if $\xi \leqq 0$.

The partial derivative W_t is continuous on $[0, 1] \times \mathbb{R}^n$, but the partial derivative W_ξ is discontinuous at all points $(\tau, 0)$ with $0 \leqq \tau \leqq 1$. Thus, the classical dynamic programming approach to this problem fails. We show how our theory applies. For (τ, ξ) with $\xi > 0$ we have that $DW(\tau, \xi; 1, h)$ exists for all $h \in Q(t, x) = [-1, 1]$ and $DW(\tau, \xi; 1, h) = 2(1 - h)[\xi + (1 - \tau)]$. Hence $\min\{DW(\tau, \xi; 1, h): h \in [-1, 1]\} = 0$ and occurs at $h = 1$. Thus $F(\tau, \xi) = 1$ if $\xi > 0$. For (τ, ξ) with $\xi < 0$ we calculate that $DW(\tau, \xi; 1, h) = -2(1 + h)[\xi - (1 - \tau)]$, so that $F(\tau, \xi) = -1$. For points (τ, ξ) with $\xi = 0$ we have that if $h \geqq 0$, then $DW(\tau, 0; 1, h) = 2(1 - \tau)(1 - h)$, while if $h \leqq 0$, then $DW(\tau, 0; 1, h) = 2(1 - \tau)(1 + h)$. Hence $F(\tau, 0) = \{+1\} \cup \{-1\}$.

The mapping F clearly satisfies property (Q) at all points of $[0, 1) \times \mathbb{R}$, except for those points on the line $\xi = 0$. We take the exceptional set $\mathscr{E}$ to be the line segment $\{(\tau, \xi): 0 \leqq \tau < 1, \xi = 0\}$. It is clear for any (τ, ξ) with $\xi \neq 0$, that the Euler polygons will all be the unique optimal trajectories from the point and will never intersect $\mathscr{E}$. For initial points on $\mathscr{E}$, the Euler polygons will either be lines of slope plus one or minus one, depending on whether $h_0 = +1$ or $h_0 = -1$. In either case the resulting trajectory only has the initial point in common with $\mathscr{E}$. Thus the function $\mathscr{F}$ clearly gives the optimal feedback.

6. Other terminal sets. The terminal sets that we have considered in this paper are $(n+1)$-dimensional with n-dimensional boundaries. In many applications the terminal sets $\mathscr{T}$ are of lower dimension. For example, in the "time optimal to the origin" problem the terminal set is the t-axis.

One treatment of such problems is to replace the problem by an approximation to the problem in which the terminal set has the structure that we have considered here. We illustrate with the time optimal to the origin problem. We take T to be very large and $\varepsilon > 0$ to be very small. We take the terminal set $\mathscr{T}$ to be the union of $[T, \infty) \times \mathbb{R}^n$ and $\mathscr{F}$, where $\mathscr{F}$ is the half-cylinder $\{(t, x): t \geqq 0, \|x\| \leqq \varepsilon\}$. The new terminal set $\mathscr{T}$ falls within the purview of our theory. If we succeed in obtaining an optimal synthesis for this problem, then we can determine those points from which we can reach $\mathscr{F}$ in minimum time t_1, where $t_1 \leqq T$. For many practical situations this model is adequate.

Another treatment of problems with lower-dimensional terminal sets that are closed is the following. The problem is first transformed into a problem with fixed terminal time, say $T = 1$. This can always be done (see, e.g., [1, p. 27]). In the transformed problem the terminal set will be some closed set $\mathscr{C}$ in the hyperplane $t = T$. The transformed problem is then replaced by a sequence of problems P_n in which the terminal set $\mathscr{T}$ is always $[T, \infty) \times \mathbb{R}^n$ and the payoffs are functions $\{g_n\}$ defined by

$$g_n(t, x) = \hat{g}(x) + n \operatorname{dist}(x, \mathscr{C})$$

where $\hat{g}$ is the payoff of the transformed problem and $\operatorname{dist}(x, \mathscr{C}) = \min\{|x - y|: y \in \mathscr{C}\}$.

Each problem P_n falls within the purview of our theory and has a value function W_n. Vinter and Mendoza [11] have shown that the values W_n converge to a function $\hat{W}$ that is the value of the transformed problem. Thus the optimal feedback syntheses of the problems P_n are approximations to the optimal synthesis for the transformed problem.

REFERENCES

[1] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[2] ———, *A Hamilton-Jacobi theory for a class of control problems*, Colloque sur la Théorie Mathématique du Contrôle Optimal, CBRM Vander, Louvain, 1970, pp. 1-23.

[3] V. G. BOLTYANSKII, *Sufficient conditions for optimality and the justification of the dynamic programming method*, SIAM J. Control Optim., 4 (1966), pp. 326-361.

[4] C. CARATHÉODORY, *Calculus of Variations and Partial Differential Equations of the First Order* (English translation by R. Dean, J. J. Brandstatter, ed.), Holden-Day, San Francisco, 1967.

[5] L. CESARI, *Optimization—Theory and Applications, Problems with Ordinary Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[6] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, Chichester, Brisbane, Toronto, Singapore, 1983.

[7] F. H. CLARK 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] H. FRANKOWSKA, *Optimal trajectories associated to a solution of the contingent Hamilton-Jacobi equation*, Appl. Math. Optim., 19 (1989), pp. 291-311.

[9] R. ISAACS, *Differential Games*, John Wiley, New York, London, Sydney, 1965.

[10] P.-L. LIONS AND P. E. SOUGANDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaacs' equations*, SIAM J. Control Optim., 23 (1985), pp. 566-583.

[11] R. B. VINTER AND L. A. MENDOZA, *Global optimality conditions for nonnormal control problems*, IMA J. Math. Control Inform., 2 (1985), pp. 241-250.

[12] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, London, Toronto, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1007-1025, September 1989

HAMILTONIAN TRAJECTORIES AND DUALITY IN THE OPTIMAL CONTROL OF LINEAR SYSTEMS WITH CONVEX COSTS*

R. T. ROCKAFELLAR†

This paper is dedicated to the memory of E. J. McShane.

Abstract. A duality theorem is proved for problems of optimal control of linear dynamical systems in continuous time subject to linear constraints and convex costs, such as penalties. Optimality conditions are stated in terms of a "minimaximum principle" in which the primal and dual control vectors satisfy a saddle point condition at almost every instant of time. This principle is shown to be equivalent to a generalized Hamiltonian differential equation in the primal and dual state variables, along with a transversality condition that likewise is in Hamiltonian form.

Key words. convex optimal control, duality, Hamiltonian trajectories, generalized problems of Bolza, calculus of variations, continuous convex programming, intertemporal convex programming

AMS(MOS) subject classifications. 49B10, 90C20, 90C05

1. Introduction. This paper focuses on optimal control problems of convex type and the special properties they enjoy, in particular, properties of duality. A fundamental problem form, intended for approximations of more complicated control situations as well as direct use in mathematical modeling, is introduced in terms of linear dynamics and linear constraints that may be represented by penalties, either finite or infinite. A duality theorem is proved and is made the basis for deriving necessary and sufficient conditions for the optimality of control functions and state trajectories. The work extends the author's recent results on continuous time problems with piecewise linear-quadratic costs [1], [2]. It ties in more generally with the theory of dual problems of Bolza in the calculus of variations, as developed earlier by methods of convex analysis in Rockafeller [3], [4]. A bridge is thereby provided to a conceptual framework dominated by a Hamiltonian function and its gradients or subgradients in the expression of optimality condition.

The chief aim, besides setting up the duality, is to demonstrate that solutions to problems in the chosen class can be characterized in two quite different, yet equivalent, ways. First, there is a "minimaximum principle," which expresses the primal and dual optimal control vectors at any time as giving a saddle point of a certain convex-concave function. Second, there is a generalized Hamiltonian differential equation in terms of primal and dual states but no direct mention of controls.

The minimaximum principle is suggestive of computational approaches that depend on generating sequences of control functions as in various algorithms of convex programming. The Hamiltonian system, on the other hand, is of interest in that it can be solved like an ordinary differential equation from any choice of initial primal and dual states. While this may or may not be a practical tool in calculating optimal trajectories, it reveals important information about such trajectories, for example, that under our assumptions they can be realized by optimal control functions that are

* Received by the editors December 2, 1988; accepted for publication December 16, 1988. This work was supported in part by grants from the National Science Foundation and the Air Force Office of Scientific Research at the University of Washington, Seattle, Washington.

† Department of Mathematics, University of Washington, GN-50, Seattle, Washington, 98105.

essentially bounded. Knowledge of the Hamiltonian function is crucial also to the prospects of applying Hamilton-Jacobi theory in its latest forms to convex problems of optimal control.

The model problem we start from is not the broadest possible problem that would fit under the heading of convex optimal control. It is selected, rather, to yield strong results while still encompassing a wide spectrum of applications. The details of structure are designed to facilitate dualization.

To help keep formulas compact and readable, we write x_t and u_t as the state and control vectors at time t instead of $x(t)$ and $u(t)$. These vectors belong to $\mathbb{R}^n$ and $\mathbb{R}^k$, respectively. We also make use of an auxiliary control vector $u_e \in \mathbb{R}^{k_e}$, which affects endpoint costs and constraints; the subscript e is utilized also to designate data elements connected with endpoints. (See [1] for a discussion of the modeling possibilities with endpoint controls.) Inner products of vectors in $\mathbb{R}^n$ and $\mathbb{R}^k$ will be expressed in the notation $\langle \cdot, \cdot \rangle$ and the Euclidean norm by $|\cdot|$.

We denote by $\mathcal{U}$ the space of all control elements u consisting of a choice of vector u_e and an *essentially bounded*, measurable function $t \mapsto u_t$ defined over the interval $[t_0, t_1]$, which is fixed throughout the paper. We handle $\mathcal{U}$ as a Banach space in the norm $\|u\| = \max\{|u_e|, \operatorname{ess\,sup}_t |u_t|\}$. Each $u \in \mathcal{U}$ determines a state trajectory $x: t \mapsto x_t \in \mathbb{R}^n$, which is Lipschitz continuous over $[t_0, t_1]$. The time derivative of x_t, which exists for almost every t, is denoted by $\dot{x}_t$. The space of all such Lipschitz continuous arcs x in $\mathbb{R}^n$ is denoted by $\mathcal{A}^\infty = \mathcal{A}_n^\infty[t_0, t_1]$. This is a Banach space in the norm $\|x\|_\infty = \max\{|x_{t_0}|, \operatorname{ess\,sup}_t \dot{x}_t\}$. (The superscript ∞ is a reminder that the derivative function $t \mapsto \dot{x}_t$ belongs to $\mathcal{L}_n^\infty[t_0, t_1]$.)

The control problem we address takes the following form:

Minimize the functional

$$(\mathcal{P}) \qquad \begin{aligned} F(u) = &\int_{t_0}^{t_1} [\langle p_t, u_t \rangle + \varphi_t(u_t) + \psi_t(q_t - C_t x_t - D_t u_t) - \langle c_t, x_t \rangle]\, dt \\ &+ [\langle p_e, u_e \rangle + \varphi_e(u_e) + \psi_e(q_e - C_e x_{t_1} - D_e u_e) - \langle c_e, x_{t_1} \rangle] \end{aligned}$$

over $u \in \mathcal{U}$, where x is determined from u by

$$\dot{x}_t = A_t x_t + B_t u_t + b_t \text{ a.e.}, \quad x_{t_0} = B_e u_e + b_e.$$

Here φ_t and φ_e are functions on $\mathbb{R}^k$ and $\mathbb{R}^{k_e}$, while ψ_t and ψ_e are functions on certain spaces $\mathbb{R}^l$ and $\mathbb{R}^{l_e}$. The dimensions of the various vectors and matrices in $(\mathcal{P})$ are of course completely determined by the dimensions of these spaces. In general we assume the following:

(A1) φ_t, φ_e, ψ_t, ψ_e, are lower semicontinuous, proper, convex functions.

(A2) φ_t and ψ_t depend epi-continuously on $t \in [t_0, t_1]$.

(A3) A_t, B_t, b_t, C_t, c_t, D_t, p_t, q_t, depend continuously on $t \in [t_0, t_1]$.

By (A3) we are assured, in particular, that each choice of $u \in \mathcal{U}$ gives rise to a unique trajectory x, which belongs to the space $\mathcal{A}^\infty$ because the function $t \mapsto A_t x_t + B_t u_t + b_t$ is essentially bounded. The mapping $u \mapsto x$ is continuous. The properness in (A1) asserts that the functions φ_t, φ_e, ψ_t, ψ_e, do not take on the value $-\infty$, although they might in some cases take on ∞ as long as they do not have this value everywhere. The role of ∞ is to provide an infinite penalty for certain constraint violations; more about this will follow.

Assumption (A2) means that the epigraphs sets epi φ_t and epi ψ_t, which are closed convex subsets of $\mathbb{R}^{k+1}$ and $\mathbb{R}^{l+1}$, depend continuously on t in the sense of set

convergence. This form of continuity has been studied by many authors in recent years (see Salinetti and Wets [5], Wets [6], for properties and references). As a special case, of course, epi-continuity is present when φ_t and ψ_t do not actually vary with t.

PROPOSITION 1.1. *Under* (A1)-(A3), *the functional F in problem* $(\mathcal{P})$ *is well defined on the Banach space* $\mathcal{U}$ *with values in* $(-\infty, \infty]$. *Furthermore, F is convex and lower semicontinuous.*

Proof. The epi-continuity of $t \mapsto \operatorname{epi} \varphi_t$ in (A2) entails that the function $(t, w) \mapsto \varphi_t(w)$ is lower semicontinuous on $[t_0, t_1] \times \mathbb{R}^k$. Therefore, this function is definitely a normal integrand in the sense of [7] and is bounded below on $[t_0, t_1] \times W$ for every bounded set $W \subset \mathbb{R}^k$. It follows that $\varphi_t(u_t)$ is measurable in t when u_t is measurable in t, and it is essentially bounded from below when u_t is essentially bounded in t. For any $u \in \mathcal{U}$, then, the integral of $\varphi_t(u_t)$ has a well-defined value in $(-\infty, \infty]$. Similar properties hold for ψ_t. Since (A3) implies $q_t - C_t x_t - D_t u_t$ is a bounded measurable function of t when u_t is such a function of t (here we note that x_t, as determined by the dynamics, is continuous in t), we conclude that the integral of $\psi_t(q_t - c_t x_t - D_t u_t)$ likewise has a well-defined value in $(-\infty, \infty]$ for any $u \in \mathcal{U}$. Thus $F(u)$ is well defined on $\mathcal{U}$ with values in $(-\infty, \infty]$. The convexity of F follows, obviously, from the convexity in (A1) and the fact that the dynamical mapping $u \mapsto x$ is affine. Lower semicontinuity in the norm topology of $\mathcal{U}$ follows from the lower semicontinuity in (A1) and continuity in (A3), as well as the continuity of $u \mapsto x$, by Fatou's Lemma (cf. [7]). □

Problem $(\mathcal{P})$ may involve *implicit constraints* beyond the ones already mentioned, due to the possibility of ∞ values. Recall that the effective domain dom F consists of the elements $u \in \mathcal{U}$ such that $F(u) < \infty$; similarly for dom φ_t and dom φ_e. Minimizing F over $\mathcal{U}$ is the same as minimizing F over dom F. Obviously the condition $u \in \operatorname{dom} F$ requires u to belong to the set

$$U := \left\{ u \in \mathcal{U} \;\middle|\; \int_{t_0}^{t_1} \varphi_t(u_t)\, dt < \infty \text{ and } \varphi_e(u_e) < \infty \right\} \tag{1.1}$$

and satisfy

$$u_t \in U_t \text{ a.e. and } u_e \in U_e, \text{ where } U_t := \operatorname{dom} \varphi_t \text{ and } U_e := \operatorname{dom} \varphi_e, \tag{1.2}$$

$$q_t - C_t x_t - D_t u_t \in R_t \text{ a.e. and } q_e - C_e x_{t_1} - D_e u_e \in R_e, \quad \text{where } R_t := \operatorname{dom} \psi_t \text{ and } R_e := \operatorname{dom} \psi_e. \tag{1.3}$$

The control problem *dual* to $(\mathcal{P})$ involves dual states $y_t \in \mathbb{R}^n$ and dual controls $v_t \in \mathbb{R}^l$ and $v_e \in \mathbb{R}^{l_e}$. Let us denote by $\mathcal{V}$ the set of control elements v consisting of a choice of v_e and an essentially bounded, measurable function $t \mapsto v_t$. This is a Banach space in the same way as described above for $\mathcal{U}$. The dual problem takes the following form:

Maximize the function

$$\begin{aligned} G(v) = \int_{t_0}^{t_1} & [\langle q_t, v_t \rangle - \psi_t^*(v_t) - \varphi_t^*(B_t^* y_t + D_t^* v_t - p_t) - \langle b_t, y_t \rangle]\, dt \\ & + [\langle q_e, v_e \rangle - \psi_e^*(v_e) - \varphi_e^*(B_e^* y_{t_0} + D_e^* v_e - p_e) - \langle b_e, y_{t_0} \rangle] \end{aligned} \tag{2}$$

over $v \in \mathcal{V}$, where y is determined from v by

$$-\dot{y}_t = A_t^* y_t + C_t^* v_t + c_t \text{ a.e.}, \quad y_{t_1} = C_e^* v_e + c_e.$$

The asterisk on a matrix denotes transpose, but on a convex function it indicates the conjugate function (Legendre-Fenchel transform) in the sense of convex analysis [8].

Note that the dual dynamical system goes backward in time and uniquely determines a Lipschitz continuous trajectory $y \in \mathscr{A}^\infty$ for each $v \in \mathscr{V}$.

Just as (A3) is preserved in passing to transposes, assumptions (A1) and (A2) imply the corresponding properties for the conjugate functions:

(A1*) $\varphi_t^*, \varphi_e^*, \psi_t^*, \psi_e^*$, are lower semicontinuous, proper, convex functions.

(A2*) φ_t^* and ψ_t^* depend epi-continuously on $t \in [t_0, t_1]$.

The equivalence between (A2) and (A2*) follows from Wijsman's Theorem [9] on the continuity of the Legendre-Fenchel transform with respect to epi-convergence. Therefore, we immediately get the version of Proposition 1.1 that applies to the dual problem.

PROPOSITION 1.2. *Under* (A1)-(A3), *the functional G in problem* ($\mathscr{Q}$) *is well defined on the Banach space* $\mathscr{V}$ *with values in* $[-\infty, \infty)$. *Furthermore, G is concave and upper semicontinuous.*

Implicit in ($\mathscr{Q}$) are the constraints that v should belong to the set

$$V := \left\{ v \in \mathscr{V} \;\middle|\; \int_{t_0}^{t_1} \psi_t^*(v_t)\, dt < \infty \text{ and } \psi_e^*(v_e) < \infty \right\} \tag{1.4}$$

and satisfy

$$v_t \in V_t \text{ a.e. and } v_e \in V_e, \text{ where } V_t := \operatorname{dom} \psi_t^* \text{ and } V_e := \operatorname{dom} \psi_e^*, \tag{1.5}$$

$$B_t^* y_t + D_t^* v_t - p_t \in S_t \text{ a.e. and } B_e^* y_{t_0} + D_e^* v_e - p_e \in S_e, \quad \text{where } S_t := \operatorname{dom} \varphi_t^* \text{ and } S_e := \operatorname{dom} \varphi_e^*. \tag{1.6}$$

The special case of these primal and dual problems that was treated in [1] and [2] as *extended linear-quadratic* optimal control is obtained by taking

$$\begin{aligned} &\varphi_t(u_t) = \tfrac{1}{2}\langle u_t, P_t u_t\rangle \quad \text{for } u_t \in U_t, \quad \varphi_t(u_t) = \infty \quad \text{for } u_t \notin U_t, \\ &\varphi_e(u_e) = \tfrac{1}{2}\langle u_e, P_e u_e\rangle \quad \text{for } u_e \in U_e, \quad \varphi_e(u_e) = \infty \quad \text{for } u_e \notin U_e, \\ &\psi_t^*(v_t) = \tfrac{1}{2}\langle v_t, Q_t v_t\rangle \quad \text{for } v_t \in V_t, \quad \psi_t^*(v_t) = \infty \quad \text{for } v_t \notin V_t, \\ &\psi_e^*(v_e) = \tfrac{1}{2}\langle v_e, Q_e v_e\rangle \quad \text{for } v_e \in V_e, \quad \psi_e^*(v_e) = \infty \quad \text{for } v_e \notin V_e, \end{aligned} \tag{1.7}$$

for *polyhedral* sets U_t, U_e, V_t, V_e, and *positive semidefinite* symmetric matrices P_t, P_e, Q_t, Q_e. The philosophy behind this is fully explained in [1] and will not be repeated here, except to say that the functions ψ_t, ψ_e, φ_t^*, φ_e^*, are then piecewise linear-quadratic and yield a version of linear-quadratic optimal control in which piecewise linear-quadratic penalty terms may be present and are readily dualized.

In general, the terms involving ψ_t and ψ_e in ($\mathscr{P}$) may be viewed as *monitoring* the vectors $s_t = q_t - C_t x_t - D_t u_t$ and $s_e = q_e - C_e x_{t_1} - D_e u_e$. A simple example would be the one where ψ_t vanishes on a certain set K but has the value ∞ outside of K. Then the ψ_t term expresses through infinite penalties the condition that $s_t \in K$ almost everywhere. This condition might represent a system of equations or inequalities. Instead ψ_t could have finite, positive values outside of K, and then we would have a finite penalty representation of such a constraint system. Similarly, ψ_e could play this role for constraints on the endpoint x_{t_1}, while φ_t^* and φ_e^* could have such interpretations in the dual problem. Many examples are worked out in [1].

Our strongest results will eventually call for a further assumption:

(A4) φ_t and φ_e are coercive, while ψ_t and ψ_e are everywhere finite.

Coercivity of φ_t means that $\lim_{|w|\to\infty} \varphi_t(w)/|w| = \infty$, which is true in particular when the control set U_t in (1.2) is bounded; similarly for φ_e and U_e. It is known from convex

analysis [8, § 13] that φ_t and φ_e are coercive if and only if the conjugate functions φ_t^* and φ_e^* are finite everywhere. Likewise, ψ_t and ψ_e are finite everywhere if and only if ψ_t^* and ψ_e^* are coercive. Thus (A4), like the earlier assumptions, has an equivalent dual form:

(A4*) ψ_t^* and ψ_e^* are coercive, while φ_t^* and φ_e^* are everywhere finite.

The interpretation of (A4), then, is that *there are effectively no exact implicit constraints of type* (1.3) *and* (1.6) *in the primal and dual problems.* In other words, this additional assumption corresponds to the situation where all the monitoring of $q_t - C_t x_t - D_t u_t$ and $q_e - C_e x_{t_1} - D_e u_e$ in the primal problem and of $B_t^* y_t + D_t^* v_t - p_t$ and $B_e^* y_{t_0} + D_e^* v_e - p_e$ in the dual problem proceeds with *finite* values: no infinite penalties. Such a property may naturally be present in a given application, or it may be achieved as a mode of approximation for a problem one is really interested in. Anyway, we may argue that it is vital for the development of computational methods for problems like $(\mathcal{P})$ and $(\mathcal{Q})$. Conditions on x or on x and u jointly that are modeled as exact constraints can lead to serious numerical complications, whereas such conditions on u alone, as in (1.2), present relatively little difficulty. See [1] for more on this issue.

PROPOSITION 1.3. *Under* (A4), *the epi-continuity assumption* (A2) *is equivalent to having* $\varphi_t^*(r)$ *and* $\psi_t(s)$ *be continuous in* $t \in [t_0, t_1]$ *for each* $r \in \mathbb{R}^k$ *and* $s \in \mathbb{R}^l$. *Then in fact* $\varphi_t^*(r)$ *is continuous with respect to* (t, r), *and* $\psi_t(s)$ *is continuous with respect to* (t, s).

Proof. For finite convex functions, epi-continuity with respect to t is equivalent to pointwise continuity with respect to t; (see Salinetti and Wets [5, Cors. 4, 5]). Furthermore, finite convex functions whose values depend continuously on t are jointly continuous in t and their other variables [8, Thm. 10.7].

PROPOSITION 1.4. *Under assumptions* (A1)-(A3), *the sets* U *and* V *in* (1.1) *and* (1.4) *are convex and nonempty. When* (A4) *holds too,* U *is identical to the set of feasible controls for* $(\mathcal{P})$, *i.e., the elements* $u \in \mathcal{U}$ *for which* $F(u)$ *is finite, and likewise* V *is the set of feasible controls for* $(\mathcal{Q})$. *In particular, feasible controls do exist, then, for both problems.*

Proof. The convexity of U and V is obvious from their definitions by the convexity in (A1). Clearly $F(u) = \infty$ when $u \notin U$, and $G(v) = -\infty$ for $v \notin V$. According to (A2), the multifunction $t \mapsto \operatorname{epi} \varphi_t$, whose values are nonempty closed convex sets by (A1), is continuous. For such a multifunction the continuous selection theorem of Michael [10] applies: it is possible to choose $(u_t, \alpha_t) \in \operatorname{epi} \varphi_t$ continuously with respect to $t \in [t_0, t_1]$. Then $\varphi_t(u_t) \leqq \alpha_t$, so the integral of $\varphi_t(u_t)$ cannot be ∞ and therefore must be finite. Taking any u_e in U_e, a set which is nonempty by the properness of φ_e in (A1), we obtain a control element $u \in U$. Thus $U \neq \emptyset$. Any $u \in U$, on the other hand, makes all the terms in the formula for $F(u)$ in $(\mathcal{P})$ be finite except perhaps for the integral of $\psi_t(s_t)$, where $s_t = q_t - C_t x_t + D_t u_t$. The function $t \mapsto s_t$ is essentially bounded in t by (A3). The continuity of $(t, s) \mapsto \psi_t(s)$ asserted by Proposition 1.3 implies that the latter function is bounded on $[t_0, t_1] \times W$ for any bounded set $W \subset \mathbb{R}^l$. We thereby obtain the essential boundedness of $\psi_t(s_t)$ in t and hence the finiteness of its integral. This yields the desired conclusion in the case of $(\mathcal{P})$. The corresponding result for $(\mathcal{Q})$ follows by duality. □

2. Minimax representation. The close relationship between problems $(\mathcal{P})$ and $(\mathcal{Q})$ that leads to their being called dual to each other stems from a joint representation in terms of a minimax problem in $\mathcal{U} \times \mathcal{V}$. To give this, we introduce the functional

$$J(u, v) := \int_{t_0}^{t_1} J_t(u_t, v_t)\, dt + J_e(u_e, v_e) - j(u, v) \tag{2.1}$$

in the notation

$$
\begin{aligned}
J_t(u_t, v_t) &:= \langle p_t, u_t\rangle + \langle q_t, v_t\rangle - \langle v_t, D_t u_t\rangle + \varphi_t(u_t) - \psi_t^*(v_t),\\
J_e(u_e, v_e) &:= \langle p_e, u_e\rangle + \langle q_e, v_e\rangle - \langle v_e, D_t u_t\rangle + \varphi_e(u_e) - \psi_e^*(v_e),
\end{aligned} \tag{2.2}
$$

and with j taken to be the bi-affine functional on $\mathcal{U} \times \mathcal{V}$ that corresponds to the dynamics and is expressed in terms of the trajectories x and y associated with u and v by

$$
\begin{aligned}
j(u, v) &:= \int_{t_0}^{t_1} \langle y_t, B_t u_t + b_t\rangle\, dt + \langle y_{t_0}, B_e u_e + b_e\rangle\\
&= \int_{t_0}^{t_1} \langle x_t, C_t^* v_t + c_t\rangle\, dt + \langle x_{t_1}, C_e^* v_e + c_e\rangle.
\end{aligned} \tag{2.3}
$$

(The validity of the equation in (2.3) is proved in [1, § 6].) Because some of the terms in (2.2) can take on the value ∞ while others are $-\infty$, a convention is necessary to ensure that $J(u, v)$ is well defined. The one we follow is standard in convex analysis: $\infty - \infty = \infty$. This clarifies the meaning of $J_t(u_t, v_t)$ and $J_e(u_e, v_e)$ in all cases in (2.2):

$$
J_t(u_t, v_t) = \begin{cases} \text{finite value} & \text{when } u_t \in U_t \text{ and } v_t \in V_t,\\ -\infty & \text{when } u_t \in U_t \text{ and } v_t \notin V_t,\\ \infty & \text{when } u_t \notin U_t, \end{cases}
$$

$$
J_e(u_e, v_e) = \begin{cases} \text{finite value} & \text{when } u_e \in U_e \text{ and } v_e \in V_e,\\ -\infty & \text{when } u_e \in U_e \text{ and } v_e \notin V_e,\\ \infty & \text{when } u_e \notin U_e, \end{cases} \tag{2.4}
$$

where the sets U_t, U_e, V_t, V_e, are the effective domains in (1.2) and (1.5). The convention enters into the formula for $J(u, v)$ in resolving the integral as ∞ whenever the positive part of the integrand (which is always measurable by the argument given in the proof of Proposition 1.1) has integral ∞ while the negative part has integral $-\infty$. (This amounts to writing $J(u, v)$ with the terms $\int_{t_0}^{t_1} \varphi_t(u_t)\, dt$ and $-\int_{t_0}^{t_1} \psi_t^*(v_t)\, dt$ separated out and then invoking the convention $\infty - \infty = \infty$ in forming the overall sum. The first of these terms is unambiguously finite or ∞, as seen in Proposition 1.1, while the second is finite or $-\infty$.)

PROPOSITION 2.1. *The functional J is convex-concave on $\mathcal{U} \times \mathcal{V}$ with finite values on $U \times V$ but infinite values everywhere else. For each $v \in V$, $J(u, v)$ is lower semicontinuous in $u \in \mathcal{U}$, while for each $u \in U$, $J(u, v)$ is upper semicontinuous in $v \in \mathcal{V}$. The objective functionals F and G in $(\mathcal{P})$ and $(\mathcal{Q})$ are given by*

$$
F(u) = \inf_{u \in \mathcal{U}} J(u, v) = \inf_{u \in U} J(u, v) \quad \text{and} \quad G(v) = \sup_{v \in \mathcal{V}} J(u, v) = \sup_{v \in V} J(u, v).
$$

Proof. In view of the definitions of U and V in (1.1) and (1.4), the convention adopted in the formula for $J(u, v)$ entails, having

$$
J(u, v) = \begin{cases} \text{finite value} & \text{when } u \in U \text{ and } v \in V,\\ -\infty & \text{when } u \in U \text{ and } v \notin V,\\ \infty & \text{when } u \notin U. \end{cases} \tag{2.5}
$$

The fact that $J(u, v)$ is convex in u and concave in v relative to the product set $U \times V$ is obvious from the convexity of the functions φ_t, φ_e, ψ_t^*, ψ_e^*. The semicontinuity follows from (A1) and (A3) by Fatou's Lemma (cf. [7]).

To establish the formula asserted for $G(v)$, it suffices because of the infinities in (2.5) to prove the first equality in the case of $v \in V$. This is done by taking the first of the forms for $j(u, v)$ in (2.3) and calculating

$$\inf_{u \in \mathcal{U}} J(u, v) = \int_{t_0}^{t_1} [\langle q_t, v_t \rangle - \psi_t(v_t) - \langle y_t, v_t \rangle]\, dt + [\langle q_e, v_e \rangle - \psi_e(v_e) - \langle y_{t_0}, v_e \rangle]$$
$$+ \inf_{u \in \mathcal{U}} \left\{ \int_{t_0}^{t_1} [\langle p_t - B_t^* y_t, u_t \rangle + \varphi_t(u_t)]\, dt + [\langle p_e - B_e^* y_{t_0}, u_e \rangle + \varphi_e(u_e)] \right\}.$$

The infimum on the right equals $-\int_{t_0}^{t_1} \varphi_t^*(B_t^* y_t - p_t)\, dt - \varphi_e^*(B_e^* y_{t_0} - p_e)$ through the conjugacy formulas

$$\varphi_t^*(r) = \sup_{u \in \mathbb{R}^k} \{\langle r, u \rangle - \varphi_t(u)\} \quad \text{and} \quad \varphi_e^*(r) = \sup_{u_e \in \mathbb{R}^{k_e}} \{\langle r, u_e \rangle - \varphi_e(u_e)\}$$

and the fundamental theorem on conjugates of integral functionals, (cf. [7, Thm. 3C]). The proof of the formula for $F(u)$ follows the same pattern. (The apparent lack of symmetry in (2.5) is restored though the observation already made that only the values of J on $U \times V$ really matter.) □

THEOREM 2.2. *Under* (A1)-(A3), *the optimal values in problems* $(\mathcal{P})$ *and* $(\mathcal{Q})$ *always satisfy* $\inf(\mathcal{P}) \geqq \sup(\mathcal{Q})$. *A pair* $(\bar{u}, \bar{v})$ *furnishes a saddle point of* J *on* $\mathcal{U} \times \mathcal{V}$ *if and only if* $\bar{u}$ *is optimal for* $(\mathcal{P})$, $\bar{v}$ *is optimal for* $(\mathcal{Q})$, *and one actually has* $\inf(\mathcal{P}) = \sup(\mathcal{Q})$. *This saddle point condition is equivalent to the following, where* $\bar{x}$ *and* $\bar{y}$ *denote the primal and dual trajectories generated by* $\bar{u}$ *and* $\bar{v}$:

$$\begin{aligned} &(\bar{u}_t, \bar{v}_t) \text{ is a saddle point of } J_t(u_t, v_t) - \langle B_t^* \bar{y}_t, u_t \rangle - \langle C_t \bar{x}_t, v_t \rangle \text{ on } U_t \times V_t \quad \text{for a.e. } t, \\ &(\bar{u}_e, \bar{v}_e) \text{ is a saddle point of } J_e(u_e, v_e) - \langle B_e^* \bar{y}_{t_0}, u_e \rangle - \langle C_e \bar{x}_e, v_e \rangle \text{ on } U_e \times V_e. \end{aligned} \tag{2.6}$$

Proof. Up to the equivalence of the saddle point condition with (2.6), the assertions are well-known consequences of the relationship displayed in Proposition 2.1, where primal and dual objectives are derived as "halves" of a minimax problem. The saddle point condition has the means by definition that

$$\bar{u} \in \operatorname*{argmin}_{u \in U} J(u, \bar{v}) \quad \text{and} \quad \bar{v} \in \operatorname*{argmax}_{v \in V} J(\bar{u}, v).$$

Due to (2.5), it requires that $\bar{u} \in U$ and $\bar{v} \in V$. Then in terms of the notation

$$\begin{aligned} \bar{J}_t(u_t, v_t) &:= J_t(u_t, v_t) - \langle B_t^* \bar{y}_t, u_t \rangle - \langle C_t \bar{x}_t, v_t \rangle, \\ \bar{J}_e(u_e, v_t) &:= J_e(u_e, v_e) - \langle B_e^* \bar{y}_{t_0}, u_e \rangle - \langle C_e \bar{x}_{t_1}, v_e \rangle, \end{aligned} \tag{2.7}$$

it reduces, by the calculation in the proof of Proposition 2.1, to

$$\bar{u}_t \in \operatorname*{argmin}_{u_t \in \mathbb{R}^k} \bar{J}_t(u_t, \bar{v}_t) \quad \text{and} \quad \bar{v}_t \in \operatorname*{argmax}_{v_t \in \mathbb{R}^l} \bar{J}_t(\bar{u}_t, v_t) \quad \text{a.e.},$$

$$\bar{u}_e \in \operatorname*{argmin}_{u_e \in \mathbb{R}^{k_e}} \bar{J}_e(u_e, \bar{v}_e) \quad \text{and} \quad \bar{v}_e \in \operatorname*{argmax}_{v_e \in \mathbb{R}^{l_e}} \bar{J}_e(\bar{u}_e, v_e).$$

These relations assert that $(\bar{u}_t, \bar{v}_t)$ is a saddle point of $\bar{J}_t$ on $\mathbb{R}^k \times \mathbb{R}^l$ for almost every t and $(\bar{u}_e, \bar{v}_e)$ is a saddle point of $\bar{J}_e$ on $\mathbb{R}^{k_e} \times \mathbb{R}^{l_e}$. But $\bar{J}_t$ and $\bar{J}_e$ have the structure (2.4) relative to $U_t \times V_t$ and $U_e \times V_e$. The saddle points in question are therefore expressed equivalently with respect to $U_t \times V_t$ and $U_e \times V_e$. This is all that has to be proved. □

The saddle point conditions in (2.6) will be referred to as the *minimaximum principle* for $(\mathcal{P})$ and $(\mathcal{Q})$. This principle is *always sufficient* for optimality according to the Theorem 2.2, and it is necessary for optimality in any circumstances where we

happen to know that $\inf(\mathcal{P}) = \sup(\mathcal{Q})$ and that both problems have solutions. We shall prove in due course that assumption (A4) provides such a circumstance. Our method requires us to examine an auxiliary pair of problems in which trajectories are optimized without direct mention of controls. This will be done in the next section.

The minimaximum principle can be stated in terms of a duality between finite-dimensional optimization problems at every instant of time. With x and y as parameter vectors in $\mathbb{R}^n$, consider the problems

$$(\mathcal{P}_t(x, y)) \qquad \min_{u \in U_t} \{\langle p_t - B_t^* y, u\rangle + \varphi_t(u) + \psi_t(q_t - C_t x - D_t u)\},$$

$$(\mathcal{Q}_t(x, y)) \qquad \max_{v \in V_t} \{\langle q_t - C_t x, v\rangle - \psi_t^*(v) - \varphi_t(B_t^* y + D_t^* v - p_t)\},$$

for each $t \in [t_0, t_1]$ and also the problems

$$(\mathcal{P}_e(x, y)) \qquad \min_{u_e \in U_e} \{\langle p_e - B_e^* y, u_e\rangle + \varphi_e(u_e) + \psi_e(q_e - C_e x - D_e u_e)\},$$

$$(\mathcal{Q}_e(x, y)) \qquad \max_{v_e \in V_e} \{\langle q_e - C_e x, v_e\rangle - \psi_e^*(v_e) - \varphi_e(B_e^* y + D_e^* v_e - p_e)\}.$$

PROPOSITION 2.3. *The minimaximum principle* (2.6) *is equivalent to the following set of conditions on $\bar{u}$ and $\bar{v}$, as expressed through the corresponding trajectories $\bar{x}$ and $\bar{y}$:*

$$\bar{u}_t \text{ solves } (\mathcal{P}_t(\bar{x}_t, \bar{y}_t)),\ \bar{v}_t \text{ solves } (\mathcal{Q}_t(\bar{x}_t, \bar{y}_t)),\ \inf(\mathcal{P}_t(\bar{x}_t, \bar{y}_t)) = \sup(\mathcal{Q}_t(\bar{x}_t, \bar{y}_t)),$$

$$\bar{u}_e \text{ solves } (\mathcal{P}_e(\bar{x}_{t_1}, \bar{y}_{t_0})),\ \bar{v}_e \text{ solves } (\mathcal{Q}_e(\bar{x}_{t_1}, \bar{y}_{t_0})),\ \inf(\mathcal{P}_e(\bar{x}_{t_1}, \bar{y}_{t_0})) = \sup(\mathcal{Q}_e(\bar{x}_{t_1}, \bar{y}_{t_0})).$$

Proof. Elementary minimax theory informs us that $(\bar{u}_t, \bar{v}_t)$ has the saddle point property in (2.6) for a given t if and only if $\bar{u}_t$ minimizes over $u \in U_t$ the function

$$f_t(u) := \sup_{v \in V_t} \{J_t(u, v) - \langle B_t^* \bar{y}_t, u\rangle - \langle C_t \bar{x}_t, v\rangle\},$$

$\bar{v}_t$ maximizes over $v \in V_t$ the function

$$g_t(v) := \inf_{u \in U_t} \{J_t(u, v) - \langle B_t^* \bar{y}_t, u\rangle - \langle C_t \bar{x}_t, v\rangle\},$$

and $\inf_{U_t} f_t = \sup_{V_t} g_t$. These functions are calculated from the reciprocal conjugacy formulas:

$$\psi_t(s) = \sup_{v \in V_t} \{\langle s, v\rangle - \psi_t^*(v)\} \quad \text{and} \quad \psi_e(s_e) = \sup_{v_e \in V_e} \{\langle s_e, v_e\rangle - \psi_e^*(v_e)\}$$

to be the objectives in $(\mathcal{P}_t(\bar{x}_t, \bar{y}_t))$ and $(\mathcal{Q}_t(\bar{x}_t, \bar{y}_t))$, respectively. The assertion concerning this pair of problems is therefore valid. The one for $(\mathcal{P}_e(\bar{x}_{t_1}, \bar{y}_{t_0}))$ and $(\mathcal{Q}_e(\bar{x}_{t_1}, \bar{y}_{t_0}))$ is similarly proved. □

3. Bolza formulations. Generalized problems of Bolza in the calculus of variations concern trajectories as elements of the space $\mathcal{A}^1 = \mathcal{A}_n^1[t_0, t_1]$ consisting of all the absolutely continuous arcs x in $\mathbb{R}^n$ over $[t_0, t_1]$. (The superscript 1 refers to the fact that the function $t \mapsto \dot{x}_t$ is an element of $\mathcal{L}_n^1[t_0, t_1]$.) Such problems have the form

$$(\mathcal{P}_B) \qquad \text{Minimize } \Phi(x) := \int_{t_0}^{t_1} L_t(x_t, \dot{x}_t)\, dt + L_e(x_{t_0}, x_{t_1}) \text{ over all } x \in \mathcal{A}^1,$$

where the functions L_t and L_e on $\mathbb{R}^n \times \mathbb{R}^n$ may be extended-real-valued. In the *convex* case, where L_t and L_e are convex functions on $\mathbb{R}^n \times \mathbb{R}^n$, there is a dual problem

$$(\mathcal{Q}_B) \qquad \text{Maximize } \Psi(y) := -\int_{t_0}^{t_1} M_t(y_t, \dot{y}_t)\, dt - M_e(y_{t_0}, y_{t_1}) \text{ over all } y \in \mathcal{A}^1,$$

in which M_t and M_e are derived from L_t and L_e by

$$M_t(y_t, \dot{y}_t) = L_t^*(\dot{y}_t, y_t) \quad \text{and} \quad M_e(y_{t_0}, y_{t_1}) = L_e^*(y_{t_0}, -y_{t_1}). \tag{3.1}$$

An extensive duality theory for convex problems of Bolza was developed in [3] and [4]. We intend to apply this theory to gain insights into the relationship between the control problems $(\mathcal{P})$ and $(\mathcal{Q})$. For this purpose we choose to define

$$\begin{aligned} L_t(x, w) &= \inf_{\substack{u \in U_t \\ A_t x + B_t u + b_t = w}} \{\langle p_t, u\rangle + \varphi_t(u) + \psi_t(q_t - C_t x - D_t u) - \langle c_t, x\rangle\}, \\ L_e(x_0, x_1) &= \inf_{\substack{u_e \in U_e \\ B_e u_e + b_e = x_0}} \{\langle p_e, u_e\rangle + \varphi_e(u_e) + \psi_e(q_e - C_e x_1 - D_e u_e) - \langle c_e, x_1\rangle\}. \end{aligned} \tag{3.2}$$

(Here x and u are temporarily just dummy vectors in $\mathbb{R}^n$ and $\mathbb{R}^k$, and similarly x_0 and x_1 in $\mathbb{R}^n$.) Our work with these expressions will make use of the concept of the *recession function* associated with a lower semicontinuous, proper, convex function f on $\mathbb{R}^n$, denoted by $\operatorname{rc} f$. Many facts about such recession functions are assembled in [8, §§ 8, 13]. We mention in particular that

$$(\operatorname{rc} f)(z) = \lim_{\lambda \to \infty} [f(\bar{z} + \lambda z) - f(\bar{z})]/\lambda \quad \text{for any } \bar{z} \in \operatorname{dom} f, \tag{3.3}$$

and that coercivity of f is equivalent to $\operatorname{rc} f$ being the indicator function δ_0 of the origin, where

$$\delta_0(z) = \infty \quad \text{for } z \neq 0, \qquad \delta_0(0) = 0.$$

PROPOSITION 3.1. *Under assumptions* (A1)-(A4), *the Bolza functional* Φ *in* $(\mathcal{P}_B)$ *is well defined on* $\mathcal{A}^1$ *and is convex. The functions* L_t *and* L_e *are themselves lower semicontinuous, proper, and convex on* $\mathbb{R}^n \times \mathbb{R}^n$, *and* L_t *depends epi-continuously on* t. *The infima defining* L_t *and* L_e *are attained whenever finite, i.e., whenever the given constraints in* (3.2) *can be satisfied. The recession functions are expressed by*

$$\begin{aligned} (\operatorname{rc} L_t)(x, w) &= (\operatorname{rc} \psi_t)(-C_t x) - \langle c_t, x\rangle + \delta_0(w - A_t x), \\ (\operatorname{rc} L_e)(x_0, x_1) &= (\operatorname{rc} \psi_e)(-C_t x_1) - \langle c_e, x_1\rangle + \delta_0(x_0). \end{aligned} \tag{3.4}$$

Proof. Consider the functions

$$\begin{aligned} K_t(x, w, u) &= \langle p_t, u\rangle + \varphi_t(u) \\ &\quad + \psi_t(q_t - C_t x - D_t u) - \langle c_t, x\rangle + \delta_0(w - A_t x - B_t u - b_t), \\ K_e(x_0, x_1, u_e) &= \langle p_e, u_e\rangle + \varphi_e(u_e) \\ &\quad + \psi_e(q_e - C_e x_1 - D_e u_e) - \langle c_e, x_1\rangle + \delta_0(x_0 - B_e u_e + b_e). \end{aligned} \tag{3.5}$$

By virtue of (A1) these are lower semicontinuous, proper, convex functions on $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k$ and $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{k_e}$. The definitions given for L_t and L_e in (3.2) are equivalent to

$$L_t(x, w) = \inf_{u \in \mathbb{R}^k} K_t(x, w, u) \quad \text{and} \quad L_e(x_0, x_1) = \inf_{u_e \in \mathbb{R}^{k_e}} K_e(x_0, x_1, u_e). \tag{3.6}$$

In the language of convex analysis, therefore, L_t is the image of K_t under the projection $(x, w, u) \mapsto (x, w)$, while L_e is the image of K_e under $(x_0, x_1, u_e) \mapsto (x_0, x_1)$. We wish to apply a general theorem about such images, namely Theorem 9.2 of [8]. This involves

a condition on the recession functions of K_t and K_e, which are calculated via (3.3) and the coercivity of φ_t and φ_e in (A4) to be

$$
\begin{aligned}
(\operatorname{rc} K_t)(x, w, u) &= \langle p_t, u\rangle + \delta_0(u) + (\operatorname{rc}\psi_t)(-C_t x - D_t u) \\
&\quad -\langle c_t, x\rangle + \delta_0(w - A_t x - B_t u) \\
&= \delta_0(u) + (\operatorname{rc}\psi_t)(-C_t x) - \langle c_t, x\rangle + \delta_0(w - A_t x), \\
(\operatorname{rc} K_e)(x_0, x_1, u_e) &= \langle p_e, u_e\rangle + \delta_0(u_e) + (\operatorname{rc}\psi_e)(-C_e x_1 - D_e u_e) \\
&\quad -\langle c_e, x_1\rangle + \delta_0(x_0 - B_e u_e) \\
&= \delta_0(u_e) + (\operatorname{rc}\psi_e)(-C_e x_1) - \langle c_e, x_1\rangle + \delta_0(x_0).
\end{aligned} \tag{3.7}
$$

(We make use of the coercivity of φ_t and φ_e in replacing rc φ_t and rc φ_e by δ_0.) The fact that $(\operatorname{rc} K_t)(0, 0, u) = 0$ only for $u = 0$, and $(\operatorname{rc} K_e)(0, 0, u_e) = 0$ only for $u_e = 0$ guarantees by the theorem just cited from [8] that L_t and L_e are lower semicontinuous, proper, convex functions for which the infima in (3.6) are always attained (i.e., the ones in (3.2) are attained when the constraints can be satisfied), and that

$$(\operatorname{rc} L_t)(x, w) = \inf_{u\in\mathbb{R}^k} (\operatorname{rc} K_t)(x, w, u) \quad \text{and} \quad (\operatorname{rc} L_e)(x_0, x_1) = \inf_{u_e\in\mathbb{R}^{k_e}} (\operatorname{rc} K_e)(x_0, x_1, u_e).$$

The latter formulas are the same as those claimed in (3.4) because of the special nature of rc K_t and re K_e in (3.7).

We must verify that L_t depends epi-continuously on t. We shall do this by way of theorems of McLinden and Bergstrom [11], showing first that K_t depends epi-continuously on t. Let us write $K_t = K_t^1 + K_t^2 + K_t^3$ with

$$K_t^1(x, w, u) = \langle p_t, u\rangle + \psi_t(q_t - C_t x - D_t u) - \langle c_t, x\rangle,$$
$$K_t^2(x, w, u) = \varphi_t(u), \qquad K_t^3(x, w, u) = \delta_0(w - A_t x - B_t u).$$

The functions in this decomposition are lower semicontinuous, proper, and convex on $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k$. We argue first that each depends epi-continuously on t. This is obvious for K_t^2 because of (A2). It holds for K_t^1 because this is a finite convex function by (A4) whose values depend continuously on t (cf. Proposition 1.3). (A finite convex function depends epi-continuously on t if and only if its value at each point depends continuously on t [5, Cors. 4, 5].) In the case of K_t^3 the epi-continuity follows from Theorem 8 of [11] because the linear transformation $(x, w, u) \mapsto w - A_t x - B_t u$ depends continuously on t (by (A3)) and has all of $\mathbb{R}^n$ as its range. We deduce next from Theorem 5 of [11] that $K_t^2 + K_t^3$ depends epi-continuously on t, because the set dom $K_t^2 -$ dom K_t^3 is all of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k$ and therefore certainly contains the origin in its interior. The same theorem of [11] applied to $K_t^1 + (K_t^2 + K_t^3)$ then yields the desired epi-continuity of K_t with respect to t, since dom $K_t^1 -$ dom $(K_t^2 + K_t^3)$ too is all of $\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k$. Recalling now that L_t is the image of K_t under the projection $(x, w, u) \mapsto (x, w)$, and $(\operatorname{rc} K_t)(0, 0, u) = 0$ only for $u = 0$, we obtain from Theorem 7 of [11] that L_t depends continuously on t. This property of L_t implies in particular that $L_t(x, w)$ is lower semicontinuous with respect to (t, x, w). The integrand in the formula for the Bolza functional Φ is certainly measurable then. The functional is well defined in this case under the $\infty - \infty = \infty$ convention explained earlier. Its convexity follows from that of L_t and L_e. □

COROLLARY 3.2. *Under assumptions* (A1)-(A4), *the dual Bolza functional* Ψ *in* $(\mathcal{Q}_B)$ *is well defined on* $\mathcal{A}_n^1[t_0, t_1]$ *and is concave. The functions* M_t *and* M_e *are lower semicontinuous, proper, and convex on* $\mathbb{R}^n \times \mathbb{R}^n$, *and* M_t *depends epi-continuously on* t. *These functions satisfy the reciprocal conjugacy formulas*

$$L_t(x_t, \dot{x}_t) = M_t^*(\dot{x}_t, x_t) \quad \text{and} \quad L_e(x_{t_0}, x_{t_1}) = M_e^*(x_{t_0}, -x_{t_1}). \tag{3.8}$$

Proof. This merely invokes the basic properties of the Legendre-Fenchel transform [8, § 12], including the fact that it preserves epi-convergence of convex functions [9].

Now it will be demonstrated that the dual Bolza functional Ψ bears the same relationship to the elements of the dual control problem $(\mathcal{Q})$ as the primal Bolza functional Φ does through (3.2) to the elements of $(\mathcal{P})$. □

PROPOSITION 3.3. *Under* (A1)-(A4), *the dual functions M_t and M_e are expressed by*

$$
\begin{aligned}
-M_t(y, z) &= \sup_{\substack{v \in V_t \\ A_t^* y + C_t^* v + c_t = -z}} \{\langle q_t, v\rangle - \psi_t^*(v) - \varphi_t^*(B_t^* y + D_t^* v - p_t) - \langle b_t, y\rangle\}, \\
-M_e(y_0, y_1) &= \sup_{\substack{v_e \in V_e \\ C_e^* v_e + c_e = y_1}} \{\langle q_e, v_e\rangle - \psi_e^*(v_e) - \varphi_e^*(B_e^* y_0 + D_e^* u_e - p_e) - \langle b_e, y_0\rangle\}
\end{aligned} \tag{3.9}
$$

where the suprema are attained whenever the indicated constraints can be satisfied. The recession functions of M_t and M_e are given by

$$
\begin{aligned}
(\mathrm{rc}\, M_t)(y, z) &= (\mathrm{rc}\, \varphi_t^*)(-B_t^* y) - \langle b_t, y\rangle + \delta_0(z + A_t^* y), \\
(\mathrm{rc}\, M_e)(y_0, y_1) &= (\mathrm{rc}\, \varphi_e^*)(-B_e^* y_0) - \langle b_e, y_0\rangle + \delta_0(y_1).
\end{aligned} \tag{3.10}
$$

Proof. Starting toward the proof of the formula for M_t in (3.9), we observe that the definition of M_t in (3.1), which means

$$
M_t(y, z) = \sup_{x,w} \{\langle z, x\rangle + \langle y, w\rangle\} = L_t(x, w),
$$

can be combined with the specification of L_t in (3.2) to yield

$$
\begin{aligned}
M_t(y, z) &= \sup_{x,u} \{\langle z, x\rangle + \langle y, A_t x + B_t u + b_t\rangle - \langle p_t, u\rangle \\
&\qquad - \varphi_t(u) - \psi_t(q_t - C_t x - D_t u) + \langle c_t, x\rangle\} \\
&= \langle b_t, y\rangle - \inf_{x,u} \{f(x, u) - g(E(x, u))\},
\end{aligned} \tag{3.11}
$$

where E is the linear transformation given by $E(x, u) = C_t x + D_t u$ and f and g are the convex and concave functions given by

$$
\begin{aligned}
f(x, u) &= \varphi_t(u) - \langle z + A_t^* y + c_t, x\rangle - \langle B_t^* y + p_t, u\rangle, \\
g(s) &= -\psi_t(q_t - s).
\end{aligned}
$$

Inasmuch as g is finite everywhere by (A4), we can apply Fenchel's Duality Theorem as stated in Corollary 31.2.1 of [8] to write

$$
\inf_{x,u} \{f(x, u) - g(E(x, u))\} = \max_{v} \{g^*(v) - f^*(E^*(v))\}
$$

where the "max" indicates attainment. The adjoint linear transformation E^* takes v into the pair $(C_t^* v, D_t^* v)$. Direct calculation of the conjugate functions f^* and g^* yields

$$
\begin{aligned}
f^*(s, r) &= \delta_0(s + z + A_t^* y + c_t) + \varphi_t^*(r + B_t^* y - p_t), \\
g^*(v) &= \langle q_t, v\rangle - \psi_t^*(v).
\end{aligned}
$$

Therefore

$$
\begin{aligned}
-M_t(y, z) = &-\langle b_t, y\rangle \\
&+ \max_{v} \{\langle q_t, v\rangle - \psi_t^*(v) - \delta_0(C_t^* v + z + A_t^* y + c_t) - \varphi_t^*(D_t^* v + B_t^* y - p_t)\}.
\end{aligned}
$$

This is equivalent to the formula asserted in (3.9). The argument for M_e runs parallel. We have from (3.1) that

$$M_e(y_0, y_1) - \sup_{x_0, x_1} \{\langle y_0, x_0\rangle + \langle y_1, x_1\rangle - L_e(x_0, x_1)\},$$

and combining this with (3.2) we get

$$\begin{aligned} M_e(y_0, y_1) = \sup_{x_0, x_1} \{&\langle y_0, B_e u_e + b_e\rangle - \langle y_1, x_1\rangle - \langle p_e, u_e\rangle - \varphi_e(u_e) \\ &- \psi_e(q_e - C_e x_1 - D_e u_e) - \langle c_e, x_1\rangle\} \\ = \langle b_e, y_0\rangle - \inf_{x_1, u_e} &\{f_e(x_1, u_e) - g_e(E_e(x_1, u_e))\} \end{aligned}$$

where $E_e(x_1, u_e) = C_e x_1 + D_e u_e$ and

$$f_e(x_1, u_e) = \varphi_e(u_e) + \langle y_1, x_1\rangle - \langle B_e^* y_0 - p_e, u_e\rangle,$$

$$g_e(s_e) = -\psi_e(q_e - s_e).$$

Fenchel's Duality Theorem brings us to

$$-M_e(y_0, y_1) = -\langle b_e, y_0\rangle + \max_{v_e} \{g_e^*(v_e) - f^*(E_e^*(v_3))\},$$

where $E_e^*(v_e) = (C_e^* v_e, D_e^* v_e)$ and

$$f_e^*(s, r_e) = \delta_0(y_1) + \varphi_e^*(r_e + B_e^* y_0 - p_e),$$

$$g_e^*(v_e) = \langle q_e, v_e\rangle - \psi_e^*(v_e),$$

and this representation is equivalent to the one claimed for M_e in (3.9). Because of the symmetry between the formulas in (3.9) and (3.2), we can obtain the recession function expressions in (3.10) by appealing to Proposition 3.1 in dual form. □

These results prepare us for demonstrating that the Bolza problems $(\mathcal{P}_B)$ and $(\mathcal{Q}_B)$ are reduced representations of control problems quite close to, but somewhat broader than, $(\mathcal{P})$ and $(\mathcal{Q})$. The extended control problems, which we denote by $(\mathcal{P}')$ and $(\mathcal{Q}')$, are obtained simply by replacing $\mathcal{U}$ and $\mathcal{V}$ by the slightly larger control spaces:

$$\mathcal{U}' := \{u \mid u_e \in \mathbb{R}^{k_e},\ u_t \in \mathbb{R}^k \text{ measurable in } t \text{ with } t \mapsto B_t u_t \text{ summable}\},$$

$$\mathcal{V}' := \{v \mid v_e \in \mathbb{R}^{l_e},\ v_t \in \mathbb{R}^l \text{ measurable in } t \text{ with } t \mapsto C_t^* v_t \text{ summable}\}.$$

Thus the extended primal problem is

$(\mathcal{P}')$ minimize the functional

$$\begin{aligned} F(u) = &\int_{t_0}^{t_1} [\langle p_t, u_t\rangle + \varphi_t(u_t) + \psi_t(q_t - C_t x_t - D_t u_t) - \langle c_t, x_t\rangle]\, dt \\ &+ [\langle p_e, u_e\rangle + \varphi_e(u_e) + \psi_e(q_e - C_e x_{t_1} - D_e u_e) - \langle c_e, x_{t_1}\rangle] \end{aligned}$$

over $u \in \mathcal{U}'$, where x is determined from u by

$$\dot{x}_t = A_t x_t + B_t u_t + b_t \text{ a.e.}, \quad x_{t_0} = B_e u_e + b_e$$

while the extended dual problem is

Maximize the functional

$$(\mathcal{Q}') \qquad G(v) = \int_{t_0}^{t_1} [\langle q_t, v_t\rangle - \psi_t^*(v_t) - \varphi_t^*(B_t^* y_t + D_t^* v_t - p_t) - \langle b_t, y_t\rangle]\, dt$$
$$+ [\langle q_e, v_e\rangle - \psi_e^*(v_e) - \varphi_e^*(B_e^* y_{t_0} + D_e^* v_e - p_e) - \langle b_e, y_{t_0}\rangle]$$

over $v \in \mathcal{V}'$, where y is determined from v by

$$-\dot{y}_t = A_t^* y_t + C_t^* v_t + c_t \text{ a.e.}, \quad y_{t_1} = C_e^* v_e + c_e.$$

Note that each $u \in \mathcal{U}'$ does determine a unique trajectory $x \in \mathcal{A}^1$ in $(\mathcal{P}')$, and similarly each $v \in \mathcal{V}'$ determines a unique $y \in \mathcal{A}^1$ in $(\mathcal{Q}')$. We shall say in this situation that x and y are *realized* by the controls u and v. For the moment we think of the functionals F and G in the extended sense of $(\mathcal{P}')$ and $(\mathcal{Q}')$ as being defined with the appropriate conventions regarding infinite values, but it will emerge from further analysis that actually $F(u) > -\infty$ and $G(v) < \infty$.

PROPOSITION 3.4. *Assume* (A1)-(A4). *Then the primal problems* $(\mathcal{P}_B)$ *and* $(\mathcal{P}')$ *are equivalent to each other in the sense that*

$$\Phi(x) = \inf\{F(u) \mid u \in \mathcal{U}', x \text{ realized by } u\}, \text{ with attainment when } \Phi(x) < \infty.$$

Likewise, the dual problems $(\mathcal{Q}_B)$ *and* $(\mathcal{Q}')$ *are equivalent to each other in the sense that*

$$\Psi(y) = \sup\{G(v) \mid v \in \mathcal{V}', y \text{ realized by } v\}, \text{ with attainment when } \Psi(y) > -\infty.$$

Proof. In terms of the functions K_t and K_e in (3.5) define

$$\Upsilon(x, u) = \int_{t_0}^{t_1} K_t(x_t, \dot{x}_t, u_t)\, dt + K_e(x_{t_0}, x_{t_1}).$$

The representations (3.6) lead to

$$(3.12) \qquad \Phi(x) = \min\{\Upsilon(x, u) \mid u_e \in \mathbb{R}^{k_e},\ u_t \in \mathbb{R}^k \text{ measurable in } t\}.$$

This is justified by the fundamental result in [12, p. 316] on control formulations versus Bolza formulations. (The inf-boundedness condition in the hypothesis of that result is fulfilled because of the recession function property of K_t established in (3.7).) Formula (3.12) is equivalent to the assertion made in the present theorem about the primal problems. Symmetry yields the corresponding fact about the dual problems. □

4. Hamiltonian functions and duality. Further progress in applying the theory of Bolza problems to the original control problems $(\mathcal{P})$ and $(\mathcal{Q})$ will depend on a study of the *Hamiltonian* function for problems $(\mathcal{P}_B)$ and $(\mathcal{Q}_B)$, which in general is defined on $\mathbb{R}^n \times \mathbb{R}^n$ by

$$(4.1) \qquad H_t(x, y) = \sup_{w \in \mathbb{R}^n} \{\langle y, w\rangle - L_t(x, w)\}.$$

PROPOSITION 4.1. *Under assumptions* (A1)-(A4), *the Hamiltonian* $H_t(x, y)$ *is finite everywhere, concave in* $x \in \mathbb{R}^n$, *convex in* $y \in \mathbb{R}^n$, *and continuous in* (t, x, y).

Proof. The fact that $H_t(x, y)$ is concave in x and convex in y follows simply from the convexity of $L_t(x, w)$ in (x, w), as in the theory of convex problems of Bolza more generally. The defining equation (4.1) says that $H_t(x, \cdot)$ is the function conjugate to $L_t(x, \cdot)$. For each choice of t and x, $L_t(x, \cdot)$ is not only lower semicontinuous and convex but proper on $\mathbb{R}^n$. This is evident from (3.2) and the finiteness of ψ_t assumed in (A4). Moreover, the recession function of $L_t(x, \cdot)$ is (rc L_t)$(0, \cdot)$ on the general basis of (3.3), and the formula in Proposition 3.1 shows (rc L_t)$(0, \cdot)$ to be δ_0. Thus $L_t(x, \cdot)$ is coercive, so that its conjugate must be finite everywhere. In other words, $H_t(x, y)$ must be finite for all (t, x, y).

We claim next that for fixed x, $L_t(x,\cdot)$ depends epi-continuously on t. This is equivalent to the assertion that the function $(z,w)\mapsto(L_t+f)(z,w)$ depends epi-continuously on t when for fixed x we define $f(z,w)=\delta_0(z-x)$. Such epi-continuity is justified by Theorem 5 of [11], because $\operatorname{dom}L_t-\operatorname{dom}f$ is all of $\mathbb{R}^n\times\mathbb{R}^n$. The Legendre-Fenchel transform preserves epi-convergence [9], so in passing to the conjugate function $H_t(x,\cdot)$ of $L_t(x,\cdot)$ we have $H_t(x,\cdot)$ depending epi-continuously on t. Because $H_t(x,\cdot)$ is finite everywhere, its epi-continuity with respect to t is the same as the continuity of $H_t(x,y)$ in t for fixed (x,y) [5, Cors. 4, 5]. This implies the continuity of $H_t(x,y)$ in (t,x,y) by [8, Thm. 35.4], due to the concavity-convexity. □

THEOREM 4.2. *Assumptions* (A1)-(A4) *guarantee that the Bolza problems* $(\mathcal{P}_B)$ *and* $(\mathcal{Q}_B)$ *both have solutions, and the same for the extended control problems* $(\mathcal{P}')$ *and* $(\mathcal{Q}')$. *Moreover,*

$$-\infty<\inf(\mathcal{P}')=\inf(\mathcal{P}_B)=\sup(\mathcal{Q}_B)=\sup(\mathcal{Q}')<\infty.$$

Proof. Only the part concerning the Bolza problems needs to be dealt with, because the rest will then follow immediately from Proposition 3.4. We shall apply the main results of the duality theory for Bolza problems in [4, Thm. 1, Thm. 3 and its Cor. 1]. The background for this application is the finiteness of the Hamiltonian as proved in Proposition 4.1, which guarantees by the corollary on p. 17 of [4] that certain basic integrability conditions, called (C_0) and (D_0) in that paper, are fulfilled. The duality results say then that we have $\infty>\inf(\mathcal{P}_B)=\sup(\mathcal{Q}_B)>\infty$ with solutions existing for both problems, provided that the following two criteria are met in terms of arcs x and y in $\mathcal{A}^1$ (this is a slightly specialized case of the results in question):

$$\int_{t_0}^{t_1}(\operatorname{rc}L_t)(x_t,\dot{x}_t)\,dt+(\operatorname{rc}L_e)(x_{t_0},x_{t_1})\leqq 0\quad\text{only for } x=0,$$

$$\int_{t_0}^{t_1}(\operatorname{rc}M_t)(y_t,\dot{y}_t)\,dt+(\operatorname{rc}M_e)(y_{t_0},y_{t_1})\leqq 0\quad\text{only for } y=0.$$

The recession function formulas provided in Propositions 3.1 and 3.3 indicate that this is indeed true in the present circumstances, because a linear ordinary differential equation has no solution starting from the origin except the 0-solution. □

The conversion of this duality and existence theorem into one for the original control problems $(\mathcal{P})$ and $(\mathcal{Q})$ will rely on the theory of optimality conditions for convex problems of Bolza as developed in [3], [4], and [13]. In addition to a generalized Hamiltonian differential equation involving subgradients of H_t, there is a transversality condition on endpoints that usually is expressed through subgradients of L_e or M_e but will now be posed in a new form. This form involves subgradients of what we shall call the *endpoint Hamiltonian*:

$$H_e(x_1,y_0):=\sup_{x_0\in\mathbb{R}^n}\{\langle x_0,y_0\rangle-L_e(x_0,x_1)\}. \tag{4.2}$$

PROPOSITION 4.3. *Under* (A1)-(A4), *the endpoint Hamiltonian* H_e *is a finite concave-convex function on* $\mathbb{R}^n\times\mathbb{R}^n$.

Proof. Definition (4.2) expresses $H_e(x_1,\cdot)$ as the function conjugate to $L_e(\cdot,x_1)$. The latter function, as seen from its definition in (3.2), is not identically ∞ for any choice of x_1 and is therefore by Proposition 3.1 a lower semicontinuous, proper, convex function on $\mathbb{R}^n$. Moreover, its recession function is $(\operatorname{rc}L_e)(\cdot,0)$, and this is δ_0 by formula (3.4) in Proposition 3.1. Hence $L_e(\cdot,x_1)$ is coercive. Consequently, its conjugate $H_e(x_1,\cdot)$ is finite everywhere [8, § 13]. This means that $H_e(x_1,y_0)$ is finite for every

choice of (x_1, y_0) in $\mathbb{R}^n \times \mathbb{R}^n$. The concavity of $H_e(x_1, y_0)$ in y_0 is a general consequence of the joint convexity of $L_e(x_0, x_1)$ in (x_0, x_1) in (4.2), cf. [8, Thm. 33.1].

Optimality conditions for the Bolza problems will be presented now in terms of subgradients of the concave-convex functions H_t and H_e. The theory of such subgradients may be found in §§ 35-37 of [8]. □

THEOREM 4.4. *For arcs $\bar{x}$ and $\bar{y}$ to be optimal for $(\mathcal{P}_B)$ and $(\mathcal{Q}_B)$ under* (A1)-(A3), *the following pair of conditions is always sufficient, and when* (A4) *holds they are also necessary*:

$$\begin{aligned} &(-\dot{\bar{y}}_t, \dot{\bar{x}}_t) \in \partial H_t(\bar{x}_t, \bar{y}_t) \quad \text{for a.e. } t \in [t_0, t_1], \\ &(\bar{y}_{t_1}, \bar{x}_{t_0}) \in \partial H_e(\bar{x}_{t_1}, \bar{y}_{t_0}). \end{aligned} \tag{4.3}$$

Proof. If the second condition in (4.3) were replaced by the usual transversality condition $(\bar{y}_{t_0}, -\bar{y}_{t_1}) \in \partial L(\bar{x}_{t_0}, x_{t_1})$, the general result would become a special case of Theorems 5 and 6 of [3], because of the equality of optimal values in Theorem 4.2. It remains only to observe that the stated conditions in terms of H_e and L_e are equivalent to each other by a general fact of subgradient theory in the case of the relationship between H_e and L_e in (4.2), namely, Theorem 37.5 of [8]. □

COROLLARY 4.5. *Suppose* (A1)-(A4) *hold. Then for an arc $\bar{x}$ to be optimal in the Bolza problem $(\mathcal{P}_B)$ it is necessary and sufficient that there exist an arc $\bar{y}$ such that the Hamiltonian conditions in* (4.3) *are satisfied. Any such arc $\bar{y}$ then solves $(\mathcal{Q}_B)$.*

Proof. This combines Theorem 4.4 with the existence assertions in Theorem 4.2. □

Generalized Hamiltonian differential equations formulated for convex problems of Bolza as in (4.3) have been studied for their own sake in [13] and also, incidentally, play a central role for Bolza problems in the nonconvex case (cf. Clarke [14]). Next we need to determine the specific form they take relative to the given data structure.

PROPOSITION 4.6. *Under* (A1)-(A4) *one has*

$$\begin{aligned} H_t(x, y) &= \langle y, A_t x\rangle + \langle b_t, y\rangle + \langle c_t, x\rangle + J_t^*(B_t^* y, C_t x), \\ H_e(x_1, y_0) &= \langle b_e, y_0\rangle + \langle c_e, x_1\rangle + J_e^*(B_e^* y_0, C_e x_1) \end{aligned} \tag{4.4}$$

where J_t^* and J_e^* are the concave-convex functions on $\mathbb{R}^n \times \mathbb{R}^n$ conjugate to J_t and J_e in (2.2) *and given by*

$$\begin{aligned} J_t^*(r, s) &= \sup_{u \in U_t} \inf_{v \in V_t} \{\langle r, u\rangle + \langle s, v\rangle - J_t(u, v)\} \\ &= \inf_{v \in V_t} \sup_{u \in U_t} \{\langle r, u\rangle + \langle s, v\rangle - J_t(u, v)\}, \\ J_e^*(r_e, s_e) &= \sup_{u_e \in U_e} \inf_{v_e \in V_e} \{\langle r_e, u_e\rangle + \langle s_e, v_e\rangle - J_e(u_e, v_e)\} \\ &= \inf_{v_e \in V_e} \sup_{u_e \in U_e} \{\langle r_e, u_e\rangle + \langle s_e, v_e\rangle - J_e(u_e, v_e)\}. \end{aligned} \tag{4.5}$$

These functions are finite everywhere, and $J_t^*(r, s)$ depends continuously on (r, s).

Proof. The conjugacy formulas

$$\begin{aligned} \psi_t(q_t - C_t x - D_t u) &= \sup_{v \in \mathbb{R}^{l_t}} \{\langle q_t - C_t x - D_t u, v\rangle - \psi_t^*(v)\} \\ &= \sup_{v \in V_t} \{\langle q_t - C_t x - D_t u, v\rangle - \psi_t^*(v)\}, \\ \psi_e(q_e - C_e x_1 - D_e u_e) &= \sup_{v_e \in \mathbb{R}^{l_e}} \{\langle q_e - C_e x_1 - D_e u_e, v_e\rangle - \psi_e^*(v_e)\} \\ &= \sup_{v_e \in V_e} \{\langle q_e - C_e x_1 - D_e u_e, v_e\rangle - \psi_e^*(v_e)\}, \end{aligned}$$

allow us to rewrite the defining formulas (3.2) for L_t and L_e in the notation (2.2) as

$$L_t(x, w) = \inf_{\substack{u \in U_t \\ A_t x + B_t u + b_t = w}} \sup_{v \in V_t} \{J_t(u, v) - \langle c_t, x\rangle - \langle C_t x, v\rangle\},$$

$$L_e(x_0, x_1) = \inf_{\substack{u \in U_e \\ B_e u + b_e = x_0}} \sup_{v_e \in V_e} \{J_e(u_e, v_e) - \langle c_e, x_1\rangle - \langle C_e x_1, v_e\rangle\}.$$

These expressions can be substituted into the definitions (4.1) of H_t and (4.2) of H_e to obtain

$$\begin{aligned} H_t(x, y) &= \sup_{u \in U_t} \{\langle y, A_t x + B_t u + b_t\rangle - \sup_{v \in V_t} \{J_t(u, v) - \langle c_t, x\rangle - \langle C_t x, v\rangle\}\} \\ &= \langle y, A_t x\rangle + \langle b_t, y\rangle + \langle c_t, x\rangle + \sup_{u \in U_t} \inf_{v \in V_t} \{\langle C_t x, v\rangle + \langle B_t^* y_t, u\rangle - J_t(u, v)\}, \\ H_e(x_1, y_0) &= \sup_{u_e \in U_e} \{\langle y_0, B_e u_e + b_e\rangle - \sup_{v_e \in V_e} \{J_e(u_e, v_e) - \langle c_e, x_1\rangle - \langle C_e x_1, v_e\rangle\}\} \\ &= \langle b_e, y_0\rangle + \langle c_t, x_1\rangle + \sup_{u \in U_e} \inf_{v \in V_e} \{\langle C_t x_1, v_e\rangle + \langle B_e^* y_0, u_e\rangle - J_e(u_e, v_e)\}. \end{aligned}$$

In the final versions of these formulas, inf and sup can be interchanged because of the coercivity of the functions φ_t, φ_e, ψ_t^*, ψ_t^*, in the definitions (2.2) of J_t and J_e and the structure (2.4). This is justified as a minimax theorem by Theorem 37.3 of [8], a result that establishes at the same time the finiteness of the expressions (4.5).

Our last task in the proof is to demonstrate that $J_t^*(r, s)$ depends continuously on (t, r, s). This could be carried out in detail with arguments like those that established the continuity of $H_t(x, y)$ in (t, x, y) in Proposition 4.1. There is a shortcut, however. The argument for H_t made no use of any particular properties of the vectors and matrices in (A3) other than their continuous dependence on t. The continuity property would therefore be present in particular if B_t and C_t were identity matrices, in which case the continuity property of H_t reduces to that of J_t. Thus $J_t(r, s)$ must be continuous with respect to (t, r, s) as claimed.

The next theorem establishes the equivalence between the Hamiltonian and minimax approaches to optimality.

THEOREM 4.7. *Under assumptions* (A1)-(A4), *the Hamiltonian optimality conditions in* (4.3) *are satisfied by a pair of arcs* $\bar{x}$ *and* $\bar{y}$ *in* $\mathscr{A}^1$ *if and only if* $\bar{x}$ *and* $\bar{y}$ *are trajectories in* $\mathscr{A}^\infty$ *realized by controls* $\bar{u} \in \mathscr{U}$ *and* $\bar{v} \in \mathscr{V}$ *that satisfy the minimaximum principle* (2.5).

Proof. This result will be developed from the following formulas for the subgradients of H_t and H_e, which are based on the representations of these functions in Proposition 4.6. Here ∂_1 and ∂_2 designate subgradients with respect to the first and second arguments of a bivariate function. We have

$$\begin{aligned} \partial_1 H_t(x, y) &= A_t^* y + c_t + C_t^* \partial_2 J_t^*(B_t^* y, C_t x), \\ \partial_2 H_t(x, y) &= A_t x + b_t + B_t \partial_1 J_t^*(B_t^* y, C_t x), \\ \partial_1 H_e(x_1, y_0) &= c_e + C_e^* \partial_2 J_e^*(B_e^* y_0, C_e x_1), \\ \partial_2 H_e(x_1, y_0) &= b_e + B_e \partial_1 J_e^*(B_e^* y_0, C_e x_1). \end{aligned}$$

These formulas are obtained by the calculus in Theorems 23.8 and 23.9 of [8] and are justified by the finiteness of the concave-convex functions J_t and J_e that was proved in Proposition 4.6. We combine these formulas with the fact that

$$\partial H_t(x, y) = \partial_1 H_t(x, y) \times \partial_2 H_t(x, y) \quad \text{and} \quad \partial H_e(x_1, y_0) = \partial_1 H_e(x_1, y_0) \times \partial_2 H_e(x_0, y_0)$$

(cf. [8, § 35]) and similarly for J_t^* and J_e^* to see that

$$\begin{aligned}\partial H_t(x, y) &= \{(A_t^* y + C_t^* v + c_t, A_t x + B_t u + b_t) | (u, v) \in \partial J_t^*(B_t^* y, C_t x)\},\\ \partial H_e(x_1, y_0) &= \{(C_e^* v_e + c_t, B_e u_e + b_e) | (u_e, v_e) \in \partial J_e^*(B_e^* y_0, C_e x_1)\}.\end{aligned} \tag{4.6}$$

A further observation is that the subdifferentials of J_t^* and J_e^* give sets of saddle points:

$$\begin{aligned}\partial J_t^*(B_t^* y, C_t x) &= \operatorname*{argminimax}_{u \in U_t, v \in V_t} \{J_t(u, v) - \langle B_t^* y, u\rangle - \langle C_t x, v\rangle\},\\ \partial J_e^*(B_e^* y_0, C_e x_1) &= \operatorname*{argminimax}_{u_e \in U_e, v_e \in V_e} \{J_e(u_e, v_e) - \langle B_e^* y_0, u_e\rangle - \langle C_e x_1, v_e\rangle\},\end{aligned} \tag{4.7}$$

which is true by conjugacy [8, Thms. 36.6, 37.5]. As far as endpoints are concerned, we have from (4.6) and (4.7) that

$$\begin{aligned}&(\bar y_{t_1}, \bar x_{t_0}) \in \partial H_e(\bar x_{t_1}, \bar y_{t_0}) \Leftrightarrow\\ &\quad \exists (\bar u_e, \bar v_e) \in \operatorname*{argminimax}_{u_e \in U_e, v_e \in V_e} \{J_e(u_e, v_e) - \langle B_e^* y_0, u_e\rangle - \langle C_e x_1, v_e\rangle\}\\ &\qquad\qquad \text{with } \bar x_{t_0} = B_e \bar u_e + b_e, \quad \bar y_{t_1} = C_e^* \bar v_e + c_e.\end{aligned} \tag{4.8}$$

Similarly, for any t it is true that

$$\begin{aligned}&(-\dot{\bar y}_t, \dot{\bar x}_t) \in \partial H_t(\bar x_t, \bar y_t) \Leftrightarrow \exists (\bar u_t, \bar v_t) \in \operatorname*{argminimax}_{u \in U_t, v \in V_t} \{J_t(u, v) - \langle B_t^* y, u\rangle - \langle C_t x, v\rangle\}\\ &\qquad\qquad \text{with } \dot{\bar x}_t = A_t \bar x_t + B_t \bar u_t + b_t, \; -\dot{\bar y}_t = A_t^* \bar y_t + C_t^* \bar v_t + c_t.\end{aligned} \tag{4.9}$$

If the minimaximum principle (2.6) holds for some choice of $\bar u \in \mathcal{U}$ and $\bar v \in \mathcal{V}$, and $\bar x$ and $\bar y$ are the corresponding state trajectories in the control problems $(\mathcal{P})$ and $(\mathcal{Q})$, we do have (4.7) and (4.8), the latter for almost every t. Then, according to the formulas we have arrived at, $\bar x$ and $\bar y$ satisfy the Hamiltonian conditions (4.3) and belong to $\mathcal{A}^\infty$ instead of just $\mathcal{A}^1$.

Conversely, suppose $\bar x$ and $\bar y$ are elements of $\mathcal{A}^1$ for which the Hamiltonian conditions (2.6) hold. Certainly then there is a choice of $\bar u_e$ and $\bar v_e$ for which (4.8) holds. We know further that for almost every t we can find $\bar u_t$ and $\bar v_t$ satisfying (4.9). It must be shown that these vectors can be chosen in such a way that the functions $t \mapsto u_t$ and $t \mapsto v_t$ are *measurable and essentially bounded.* Inasmuch as J_t^* is finite everywhere, the subgradient set $\partial J_t^*(r, s)$ is always nonempty and compact (by [8, Thm. 23.4] as applied to the separate arguments). The continuity of $J_t^*(r, s)$ in t implies further that the multifunction $(t, r, s) \mapsto \partial J_t^*(r, s)$ is locally bounded and of closed graph [8, Thm. 24.5]. Therefore the multifunction $t \mapsto \partial J_t^*(\bar x_t, \bar y_t)$ is locally bounded and of closed graph, as well as nonempty-valued. A measurable selection then exists (cf. [7, Cor. 1C]) and must be essentially bounded. This selection is in the form of a function $t \mapsto (\bar u_t, \bar v_t)$ with exactly the properties we need.

We have arrived at the main duality theorem in this paper.

THEOREM 4.8. *Assumptions* (A1)-(A4) *guarantee that the primal and dual control problems* $(\mathcal{P})$ *and* $(\mathcal{Q})$ *are both solvable and satisfy*

$$-\infty < \inf(\mathcal{P}) = \inf(\mathcal{P}_B) = \sup(\mathcal{Q}_B) = \sup(\mathcal{Q}) < \infty.$$

An arc $\bar x$ *solves* $(\mathcal{P}_B)$ *if and only if it is realized by a control* $\bar u$ *that is optimal in* $(\mathcal{P})$, *and an arc* $\bar y$ *solves* $(\mathcal{Q}_B)$ *if and only if it is realized by a control* $\bar v$ *that is optimal in* $(\mathcal{P})$.

Proof. Mostly this is a corollary of Theorems 4.2, 4.4, and 4.7, combined. In general we would of course have $\inf(\mathcal{P}) \geqq \inf(\mathcal{P}')$ and $\sup(\mathcal{Q}) \leqq \inf(\mathcal{Q}')$. To obtain the full result, we need only show that when $\bar u \in \mathcal{U}$ and $\bar v \in \mathcal{V}$ satisfy the minimaximum

principle, then $F(\bar{u}) = G(\bar{v})$. The minimaximum principle implies by Proposition 2.3 that

$$\langle p_t - B_t^* \bar{y}_t, \bar{u}_t \rangle + \varphi_t(\bar{u}_t) + \psi_t(q_t - C_t \bar{x}_t - D_t \bar{u}_t)$$
$$= \langle q_t - C_t \bar{x}_t, \bar{v}_t \rangle - \psi_t^*(\bar{u}_t) - \varphi_t^*(B_t^* \bar{y}_t + D_t^* \bar{v}_t - p_t)$$

for almost every t and

$$\langle p_e - B_e^* \bar{y}_{t_0}, \bar{u}_e \rangle + \varphi_e(\bar{u}_e) + \psi_e(q_e - C_e \bar{x}_{t_1} - D_e \bar{u}_e)$$
$$= \langle q_e - C_e \bar{x}_{t_1}, \bar{v}_e \rangle - \psi_e^*(\bar{u}_e) - \varphi_e^*(B_e^* \bar{y}_{t_0} + D_e^* \bar{v}_e - p_e).$$

Integrating the first equation with respect to $t \in [t_0, t_1]$ and adding it to the second equation, we obtain

$$F(\bar{u}) - \int_{t_0}^{t_1} \langle B_t^* \bar{y}_t, \bar{u}_t \rangle \, dt + \int_{t_0}^{t_1} \langle c_t, \bar{x}_t \rangle \, dt - \langle B_e^* \bar{y}_{t_0}, \bar{u}_e \rangle + \langle c_e, \bar{x}_{t_1} \rangle$$
$$= G(\bar{v}) - \int_{t_0}^{t_1} \langle C_t \bar{x}_t, \bar{v}_t \rangle \, dt + \int_{t_0}^{t_1} \langle b_t, \bar{y}_t \rangle \, dt - \langle C_e \bar{x}_{t_1}, \bar{v}_e \rangle + \langle b_e, \bar{y}_{t_0} \rangle.$$

This reduces to $F(\bar{u}) = G(\bar{v})$ because of the identity in (2.3).

The methods used to obtain this result have relied heavily on assumption (A4), which as we have seen corresponds to the absence of any exactly modeled constraints on the primal or dual states. A more general theory in which state constraints are present might well be possible, but results for convex problems of Bolza in that case in [15]-[17], indicate it would require consideration of impulse controls and trajectories with jumps.

REFERENCES

[1] R. T. ROCKAFELLAR, *Linear-quadratic programming and optimal control*, SIAM J. Control Optim., 25 (1987), pp. 781-814.

[2] ———, *On the essential boundedness of solutions to problems in piecewise linear-quadratic optimal control*, in Analyse Mathématique et Applications, F. Murat and O. Pironneau, eds., Gauthier-Villars, Paris, 1988, pp. 437-443.

[3] ———, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174-222.

[4] ———, *Existence and duality theorems for convex problems of Bolza*, Trans. Amer. Math. Soc., 139 (1971), pp. 1-40.

[5] G. SALINETTI AND R. J-B. WETS, *On the convergence of sequences of convex sets in finite dimensions*, SIAM Rev., 21 (1979), pp. 18-33.

[6] R. J-B. WETS, *Convergence of convex functions, variational inequalities and convex optimization problems*, in Variational Inequalities and Complementarity Problems, R. W. COTTLE, F. GIANNESSI, and J.-L. LIONS, eds., John Wiley, New York, 1980, pp. 376-403.

[7] 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 543, Springer-Verlag, Berlin, New York, 1976, pp. 157-207.

[8] ———, *Convex Analysis*, Princeton University Press, Princeton NJ, 1970.

[9] R. J. WIJSMAN, *Convergence of sequences of convex sets, cones, and functions*, II, Trans. Amer. Math. Soc., 123 (1966), pp. 32-45.

[10] E. MICHAEL, *Continuous selections*, I, Ann. of Math., 63 (1956), pp. 361-382.

[11] L. MCLINDEN AND R. C. BERGSTROM, *Preservation of convergence of convex sets and functions*, Trans. Amer. Math. Soc., 268 (1981), pp. 127-142.

[12] R. T. ROCKAFELLAR, *Existence theorems for general control problems of Bolza and Lagrange*, Adv. in Math., 15 (1975), pp. 312-333.

[13] R. T. ROCKAFELLAR, *Generalized Hamiltonian equations in convex problems of Lagrange*, Pacific J. Math., 32 (1970), pp. 46-63.

[14] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[15] R. T. ROCKAFELLAR, *State constraints in convex control problems of Bolza*, SIAM J. Control Optim., 10 (1972), pp. 691-715.

[16] ———, *Dual problems of Lagrange for Arcs of bounded variation*, in Calculus of Variations and Control, Academic Press, New York, 1976, pp. 155-192.

[17] ———, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in Game Theory and Math. Economics, O. Moeschlin, ed., North-Holland, 1981, pp. 339-349.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1026-1047, September 1989

008

OPTIMIZATION OF GLOBALLY CONVEX FUNCTIONS*

T. C. HU†, VICTOR KLEE‡, AND DAVID LARMAN§

This paper is dedicated to the memory of E. J. McShane.

Abstract. Convex functions have nice properties with respect to both minimization and maximization. Similar properties are established here for functions that are permitted to have bad local behavior but are globally convex in the sense that they behave "convexly" on triples of collinear points that are widely dispersed. The results illustrate a development that seems desirable in the interest of more realistic mathematical modeling: the "globalization" of important function properties. In connection with the maximization of globally convex functions over convex bodies in a given finite-dimensional normed space E, there is interest in estimating the maximum, for points c of bodies $C \subset E$, of the ratio between two measures of how close c comes to being an extreme point of C. Good estimates are obtained for the cases in which E is Euclidean or has the "max" norm.

Key words. convex, quasiconvex, maximum, minimum, optimization, extreme points, norm

AMS(MOS) subject classifications. 52A20, 52A40, 90C25, 46B20, 46C05

Introduction. In the mathematical modeling of practical optimization problems, the following assumptions are often made:

(a) The feasible region C is a convex subset of a Minkowski space E (a normed finite-dimensional real vector space);

(b) The objective function φ is convex, that is, φ is a real-valued function on C such that

$$\varphi(y) \leqq \lambda\varphi(x) + (1-\lambda)\varphi(z) \tag{1}$$

for all

$$x, z \in C \quad \text{with } x \neq z, \quad 0 < \lambda < 1, \quad y = \lambda x + (1-\lambda)z. \tag{2}$$

Because of the following well-known facts, convexity is useful in connection with both minimization and maximization:

(P1) Each local minimum of a convex function is a global minimum.

(P2) If a convex function attains a maximum, then (under mild restrictions on the domain) it does so at an extreme point of its domain.

Each of the properties (P1) and (P2) serves to narrow the search for the extreme values of a convex function, and each is the basis of algorithms for finding or approximating these values and the points of the domain C where they are attained. (There are many references for convex minimization. See [13] for a survey of approaches to convex maximization.) However, in a number of situations it seems that (a) is fully justified while (b) is dictated as much by mathematical convenience as by realism. Even when the real objective function appears to be convex when viewed globally, it is likely to exhibit small "blips" in local behavior that cause it to deviate from the

* Received by the editors November 21, 1988; accepted for publication December 13, 1988. This research was supported in part by the National Science Foundation.

† Department of Computer Science, University of California at San Diego, La Jolla, California 92093.

‡ Department of Mathematics, University of Washington, Seattle, Washington 98195.

§ Department of Mathematics, University College, London WC1E 6BT, United Kingdom.

mathematical ideal expressed in (b). Similar statements apply to other important function properties, such as linearity and monotoneity. The main purpose of the present paper is to suggest the desirability of "globalizing" various function properties that are important in optimization—that is, of formulating definitions that permit bad local behavior while preserving the function property in some global sense—and of studying the consequences of these definitions. The resulting mathematical framework may turn out to be especially appropriate for some of the many practical optimization problems in which rapid solution is more important than precise solution, so that the practical needs can be met by any algorithm that is sufficiently fast and comes sufficiently close to finding the optimum. (This is the consideration, for example, that has led to the popularity of simulated annealing as an algorithmic tool for solving problems in VLSI design and other areas [9], [16].)

It is not clear which generalizations of the notion of convex function will prove to be most useful in modeling. That can be determined only by extensive computational practice in conjunction with development of the underlying theory. However, in the previous generalizations with which we are familiar [14], [4], [3], [1], [15], local behavior is restricted in ways that exclude the sort of blips that we want to permit. To illustrate the sort of mathematical developments that we have in mind, we here formulate some notions of global convexity or global quasiconvexity that depend on a nonnegative parameter δ. They reduce to the usual notions when $\delta = 0$, but for $\delta > 0$ they do permit wild local behavior and they lead to "δ-versions" of (P1) and (P2).

Our main δ-versions of (P1) are Theorems 2.1 and 2.2, which are straightforward and easy to prove. However, the search for quantitative precision in the case of (P2) leads to some difficult problems concerning the relationship between two measures of how close a nonextreme point c of a body C comes to being extreme. (The term *body* is used to mean a finite-dimensional line-free closed convex set, so the existence of extreme points is guaranteed.) The obvious measure is $\xi(C, c)$, the minimum distance of c from C's set of extreme points, and we want to find a φ-maximizing point c for which this distance is not too large. The search for such a point turns out to involve $\mu(C, c)$, half the length of a longest segment in C that has c as its midpoint. Our main δ-version of (P2) is Theorem 4.3, which implies that if the objective function φ is δ-convex or δ-quasiconvex in an appropriate sense, and if φ attains a maximum on C, then φ attains a maximum at a point $q \in C$ such that $\xi(C, q) \leqq \delta\rho(C)$, where

$$\rho(C) = \sup\{\xi(C, c)/\mu(C, c)\text{: nonextreme } c \in C\}.$$

By way of illustration, suppose that X is a Euclidean disk of unit radius centered at a point p. When an interior point x of X is at distance r from p, it is true that $\xi(X, x) = 1 - r$, $\mu(X, x) = (1 - r^2)^{1/2}$, and $\rho(X, x) = ((1 - r)/(1 + r))^{1/2}$; hence $\rho(X) = \rho(X, p) = 1$. If Y is an equilateral triangle inscribed in X, then $\rho(Y) = \rho(Y, p) = \sqrt{3}$. However, the supremum $\rho(C)$ appears to be difficult to compute for most choices of C and of the underlying Minkowski space E, and because of ρ's role in the extension of (P2) this leads to interest in estimating the quantity

$$\rho(E) = \sup\{\rho(C)\text{: body } C \subset E\}$$

for various choices of E. It is proved here that $\rho(E) \leqq d$ for all d-dimensional E (this bound is attained), while $d - 1 \leqq \rho(E) \leqq d$ when E has the "max" norm and

$$\sqrt{d} \leqq \rho(E) \leqq \frac{d}{d-1}\sqrt{5d}$$

when E is Euclidean. Also,

$$\rho(\text{Euclidean plane}) = \sqrt{3}.$$

Our section headings are as follows: §1. Some global versions of convexity; §2. Minima; §3. Qualitative properties of $\rho(C, c)$; §4. Maxima; §5. Estimation of $\rho(E)$.

1. Some global versions of convexity. With respect to optimization, some important properties of convex functions extend to quasiconvex functions. A function φ is *quasiconvex* if its domain is a convex set and

$$\varphi(y) \leqq \max\{\varphi(x), \varphi(z)\} \tag{3}$$

whenever (2) holds. We "globalize" the notions of convexity and quasiconvexity by saying (for $\delta \geqq 0$) that a function φ is *δ-convex* (respectively, *δ-quasiconvex*) if it is real-valued, its domain is a convex set C in a normed (real) vector space, and the inequality (1) (respectively, (3)) holds for all x, y, z, and λ that satisfy both (2) and

$$\|x-y\| \geqq \delta/2 \quad \text{and} \quad \|z-y\| \geqq \delta/2. \tag{4}$$

In other words, the function φ behaves "convexly" or "quasiconvexly" on collinear triples that are sufficiently dispersed. Note that the 0-convex functions are precisely those that are convex in the usual sense; similarly for the 0-quasiconvex functions. However, as is suggested by Fig. 1, small blips are permitted in the function's local behavior when $\delta > 0$.

A function φ is *strictly δ-convex* (respectively, *strictly δ-quasiconvex*) if it is real-valued, its domain is convex, and strict inequality holds in (1) (respectively, (3)) for all x, y, z, and λ that satisfy both (2) and (4). Actually, it suffices for present

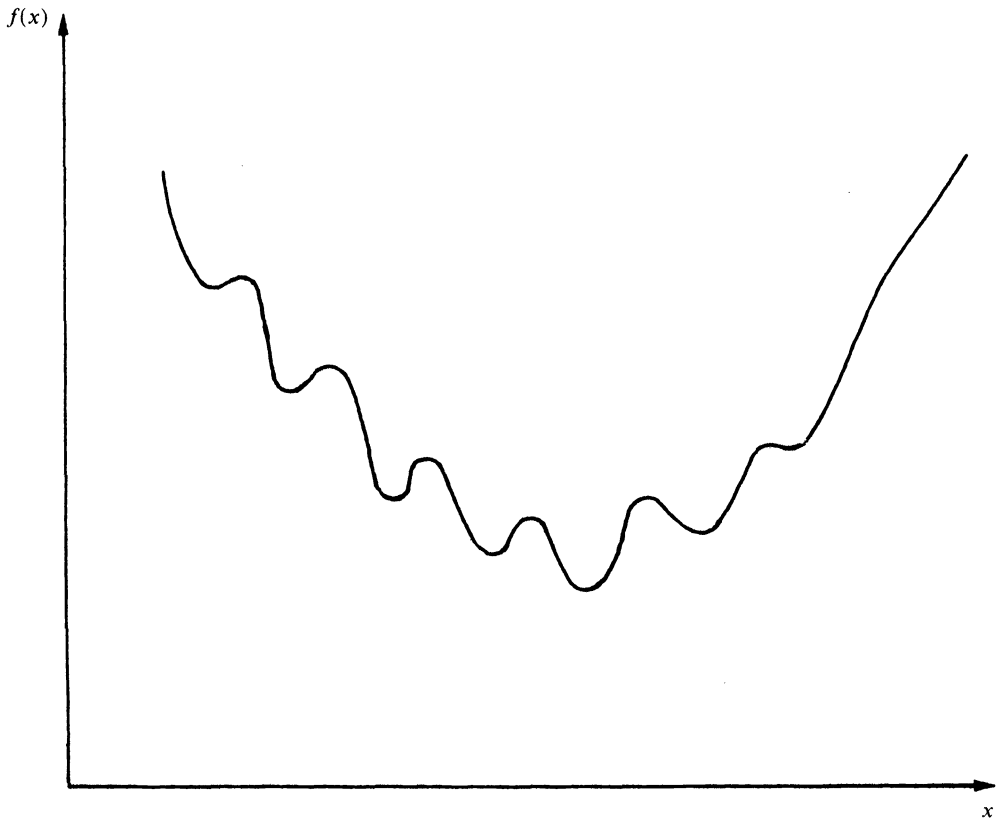

FIG. 1

purposes to require that (1) or (3) holds for all x, y, and z that satisfy (2) and (4) with $\lambda = \frac{1}{2}$. This amounts to requiring that

$$\varphi(\tfrac{1}{2}x + \tfrac{1}{2}z) \leqq \tfrac{1}{2}\varphi(x) + \tfrac{1}{2}\varphi(z) \tag{5}$$

or

$$\varphi(\tfrac{1}{2}x + \tfrac{1}{2}z) \leqq \max\{\varphi(x), \varphi(z)\} \tag{6}$$

whenever $\|x - z\| \leqq \delta$. These weak notions are called *midpoint δ-convexity and midpoint δ-quasiconvexity*, respectively, and the related strict notions are defined in the obvious way.

The following remark provides some insight into the notions just defined.

THEOREM 1.1. *For a convex subset C of a normed linear space, and for $\delta > 0$, let C^δ (respectively, C_δ) denote the set of all points $c \in C$ such that c is not the midpoint (respectively, not an endpoint) of any line segment of length δ in C. Suppose that ψ and ω are real-valued functions on C such that*

$$\psi(c) \geqq 0 \quad \text{for all } c \in C^\delta, \qquad \psi(c) = 0 \quad \text{for all } c \notin C^\delta,$$

$$\omega(c) \geqq 0 \quad \text{for all } c \in C_\delta, \qquad \omega(c) = 0 \quad \text{for all } c \notin C_\delta.$$

Then whenever a function φ on C has any of the δ-convexity properties defined above, the function $\varphi + \psi - \omega$ has the same property.

Proof. If $c \in C^\delta$, then (because of (4)) c cannot appear as the y in any of the defining inequalities (1) or (3). When c appears as the x or z in any of these inequalities, increasing the value of $\varphi(c)$ (by adding $\psi(c)$ when $\psi(c) > 0$) merely serves to strengthen the inequality. Similarly, if $c \in C_\delta$ then c cannot appear as x or z, and when c appears as y, reducing the value of $\varphi(c)$ (by subtracting $\omega(c)$ when $\omega(c) > 0$) strengthens the inequality. □

When the domain C is of diameter less than δ, every real-valued function on C is strictly δ-convex because no triple (x, y, z) of points of C satisfies condition (4). Thus it is clear that for the δ-convexity of an objective function to be useful in practice, the value of δ must be appropriately related to the geometry of the domain C. Even when the domain is in all senses large with respect to δ, the above δ-notions permit much wilder behavior (at least near extreme points of the domain) than will be encountered in practice. Nevertheless, we are able to establish sharp δ-versions of (P1) and (P2) for functions that are midpoint δ-convex or midpoint δ-quasiconvex. It does not seem that sharper conclusions could be obtained by adding local smoothing conditions such as continuity, and in any case we would not want to assume continuity because it is lacking in some important applications such as the fixed-charge problem of Hirsch and Dantzig [7].

If we wanted to augment the δ-requirement by a weak global smoothing condition to limit the wildest behavior of the function φ to the extreme points of its domain, the following assumption might be appropriate, where $]x, z[$ (respectively, $[x, z]$) denotes the open (respectively, closed) line segment whose ends are x and z.

(7) For each triple of distinct points x, y, and z of C such that $y \in]x, z[$, the segment $[x, z]$ is contained in a segment $[x', z']$ in C for which $\varphi(y) \leqq \max\{\varphi(x'), \varphi(z')\}$.

We shall see that the property (P2) follows from (7) alone, without the intervention of any δ-requirement. However, (7) has no effect on (P1).

Note that if φ (with convex domain C) satisfies condition (7), then without destroying this property the function φ may be redefined at all extreme points of C by assigning arbitrary new values that are not less than the original values. However, we do not want to rule out this aspect, because the same is true of the usual notion of convexity.

If a function φ has any of the convexity or quasiconvexity properties defined above, the same is true of each positive multiple of φ. The convexity properties are also preserved under addition of functions, but that is not true of the quasiconvexity properties or of (7). For example, let φ and ζ both have $[0, 1]$ as domain, with

$$\varphi(x) = \zeta(x) = 0 \quad \text{for } 0 < x < 1,$$
$$\varphi(0) = \zeta(1) = 1, \qquad \varphi(1) = \varphi(0) = -2.$$

Then φ and ζ are both quasiconvex and hence also satisfy condition (7). However, the function $\varphi + \zeta$ does not satisfy (7) and it is δ-quasiconvex only for $\delta < \frac{1}{2}$. This example is easily modified to produce one in which the functions are continuous.

The next two theorems deal with an example that was chosen to illustrate the extent to which midpoint δ-convexity and midpoint δ-quasiconvexity are weaker than the usual notions or convexity and quasiconvexity.

THEOREM 1.2. *Suppose that η is a positive constant, and let*

$$\varphi(t) = \eta t^2 + \cos t \quad \textit{for all } t \in \mathbb{R}.$$

Then the following three conditions are equivalent: φ is convex; φ is quasiconvex; $\eta \geqq \frac{1}{2}$.

Proof. For equivalence of the first and third condition, observe that $\varphi''(t) = 2\eta - \cos t$, whence the second derivative φ'' is everywhere nonnegative if and only if $\eta \geqq \frac{1}{2}$. To complete the proof, note that the following is a corollary of the next result: if φ is midpoint quasiconvex, then $\eta \geqq \frac{1}{2}$. □

To simplify the notation in the next statement and proof, we define

$$\psi(\delta) = \frac{1}{2}\frac{\sin^2 \delta}{\delta^2} \quad \text{for } 0 < \delta < 2\pi,$$

whence of course $\psi(\delta) \to \frac{1}{2}$ as $\delta \to 0$. It follows from l'Hôpital's rule that as $\delta \to 0$,

$$\frac{\psi(\delta)}{\frac{1}{2} - \frac{1}{6}\delta^2} \to 1$$

and hence

$$\psi(\tfrac{1}{2}\delta) \sim \tfrac{1}{2} - \tfrac{1}{24}\delta^2 \quad \text{and} \quad \psi(\tfrac{1}{4}\delta) \sim \tfrac{1}{2} - \tfrac{1}{96}\delta^2.$$

THEOREM 1.3. *Suppose that η is a positive constant, and let*

$$\varphi(t) = \eta t^2 + \cos t \quad \textit{for all } t \in \mathbb{R}.$$

Suppose that $0 < \delta < 2\pi$. Then

(i) *φ is midpoint δ-convex if and only if $\eta \geqq \psi(\frac{1}{4}\delta)$;*

(ii) *φ is midpoint δ-quasiconvex if and only if $\eta \geqq \psi(\frac{1}{2}\delta)$.*

Proof. Let us first consider (i). The function φ is midpoint δ-convex if and only if (5) holds whenever $x + \delta \leqq z$. Writing $h = z - x \geqq 0$ and setting

$$g(x, h) = 2\cos(x + \tfrac{1}{2}h) - \cos x - \cos(x + h),$$

we see that the desired condition is

$$g(x, h) \leqq \frac{\eta}{2} h^2 \quad \text{for } h \geqq \delta. \tag{8}$$

To obtain a more useful expression for $g(x, h)$, apply the trigonometric addition formulas to see that

$$g(x, h) = \cos x[2(\cos \tfrac{1}{2}h - 1) + 1 - \cos h] + \sin x(\sin h - 2 \sin \tfrac{1}{2}h),$$

then apply the half-angle formulas and regroup to obtain

$$g(x, h) = 4(\sin \tfrac{1}{4}h)^2 \cos (x + \tfrac{1}{2}h).$$

Thus in terms of the function ψ defined earlier, the desired condition (8) becomes

$$\psi(\tfrac{1}{4}\delta) \cos (x + \tfrac{1}{2}h) \leqq \eta \quad \text{for } h \geqq \delta. \tag{9}$$

Now recall that in the range $0 < \theta \leqq \pi/2$, $\sin \theta / \theta$ is a decreasing function of θ and hence

$$\frac{2}{\pi} \leqq \frac{\sin \theta}{\theta} \leqq 1.$$

Since

$$\frac{\sin \frac{1}{4}h}{\frac{1}{4}h} \leqq \frac{2}{\pi} \quad \text{when } h \geqq 2\pi,$$

it follows that the maximum value of $\psi(\frac{1}{4}h)$ for $|h| \geqq \delta$ is attained when $h = \delta$. Hence the condition

$$\eta \geqq \psi(\tfrac{1}{4}\delta) \tag{10}$$

is sufficient for (9). On the other hand, for each given h there exists an x for which $\cos (x + \frac{1}{2}h) = 1$, and hence the condition (10) is necessary as well as sufficient for the desired midpoint δ-convexity.

We turn now to (ii). The function φ is midpoint δ-quasiconvex if and only if

$$\varphi(x) \leqq \max \{\varphi(x - h), \varphi(x + h)\} \quad \text{for all } h \geqq \delta$$

(and we are concerned here only with $0 \leqq \delta \leqq \pi$). Thus the desired condition is that, with + or −, it should be true that

$$\eta x^2 + \cos x \leqq \eta(x \pm h)^2 + \cos (x \pm h),$$

or, equivalently,

$$\pm 2 \sin \frac{h}{2} \sin \left(x \pm \frac{h}{2}\right) \leqq \pm \eta h(2x \pm h). \tag{11±}$$

Now let us suppose (without loss of generality) that $x \geqq 0$ and $h \geqq 0$, and consider the inequalities

$$\frac{1}{2} \frac{\sin (h/2)}{h/2} \frac{\sin ((h/2) \pm x)}{(h/2) \pm x} \leqq \eta. \tag{12±}$$

Elementary manipulation shows that (11+) is equivalent to (12+), while (11−) is equivalent to (12−) when $2x - h < 0$ and to

$$\frac{1}{2} \frac{\sin (h/2)}{h/2} \frac{\sin ((h/2) \pm x)}{(h/2) \pm x} \geqq \eta \tag{13}$$

when $2x - h > 0$.

To deal with the case in which $x < h/2$, note that $(\sin\theta)/\theta$ is a concave function on $[-\pi, \pi]$, whence

$$\frac{\sin((h/2)-x)}{(h/2)-x}+\frac{\sin((h/2)+x)}{(h/2)+x}\leqq 2\sin\frac{h}{2},$$

and thus at least one of the summands is at most

$$\frac{\sin(h/2)}{h/2}.$$

Hence requiring that

$$\eta\geqq\frac{1}{2}\frac{\sin(\delta/2)}{\delta/2}$$

will guarantee that

$$\eta\geqq\frac{1}{2}\max_{h\geqq\delta}\left(\frac{\sin(h/2)}{h/2}\right)^2$$

and thus assure that (12±) holds. Furthermore, by taking $x=0$ and $h=\delta$ we see that this is the best possible lower bound for η when $x<h/2$.

Only the case in which $x>h/2$ remains. For this case, (12−) or (13) must be established, and we show in fact that (12−) holds. Indeed, if the number y is defined by the condition that

$$x+\frac{h}{2}=y+\frac{\delta}{2},$$

then (12−) becomes

$$\frac{\sin(y+\delta/2)}{y+\delta/2}\leqq\frac{\sin(\delta/2)}{\delta/2},$$

and this holds for all $y\geqq 0$. □

2. Minima. When $\delta>0$ and φ is a function with domain C, we say that φ attains a *δ-local minimum* (respectively, *strict δ-local minimum*) at a point $q\in C$ if $\varphi(q)\leqq\varphi(c)$ (respectively, $\varphi(q)<\varphi(c)$) for all points $c\in C\backslash\{q\}$ at distance less than δ from the point q.

THEOREM 2.1. *If $\delta>0$ and φ is a midpoint δ-convex or strictly midpoint δ-quasiconvex function that attains a δ-local minimum at a point q of its domain, then the global minimum of φ is attained at q.*

THEOREM 2.2. *If $\delta>0$ and φ is a midpoint δ-quasiconvex function that attains a δ-local strict minimum at a point q of its domain, then the strict global minimum of φ is attained at q.*

Proofs. As stated, the above theorems fail when $\delta=0$ and thus they do not cover the classical cases of convex and quasiconvex functions. To cover these cases as well, simply replace the δ-local minima by ε-local minima for some positive $\varepsilon\geqq\delta$. The proofs below are phrased in terms of this ε.

If the conclusion of Theorem 2.1 (respectively, 2.2) fails, there is a point $z_1\in C\backslash\{q\}$ such that $\varphi(z_1)<\varphi(q)$ (respectively, $\leqq\varphi(q)$). Of course, $\|z_1-q\|\geqq\varepsilon$, and hence with $z_2=\frac{1}{2}(z_1+q)$ it is true that $\|z_1-z_2\|\geqq\varepsilon/2$ and $\|q-z_2\|\geqq\varepsilon/2$. For Theorem 2.1, it follows from the midpoint δ-convexity of φ that

$$\varphi(z_2)\leqq\frac{\delta}{2}\varphi(z_1)+\frac{\delta}{2}\varphi(q)<\varphi(q),$$

or from the strict midpoint δ-quasiconvexity of φ that

$$\varphi(z_2) < \max\{\varphi(z_1), \varphi(q)\} \leqq \varphi(q).$$

For Theorem 2.2 it follows from the midpoint δ-quasiconvexity of φ that

$$\varphi(z_2) \leqq \max\{\varphi(z_1), \varphi(q)\} \leqq \varphi(q).$$

But then $\|z_2 - q\| \geqq \varepsilon$, so with $z_3 = \frac{1}{2}(z_2 + q)$ we have $\varphi(z_3) < \varphi(q)$ in Theorem 2.1 and $\varphi(z_3) \leqq \varphi(q)$ in Theorem 2.2. Continuing in this manner, we form a sequence $z_1, z_2, \cdots$ in $C\backslash\{q\}$ such that always $\varphi(z_i) < \varphi(q)$ (respectively, $\leqq \varphi(q)$). Since the sequence converges to q, we have reached a contradiction that completes the proof. □

Theorems 2.1 and 2.2 are sharp in the following two senses:

(a) In Theorem 2.1, the conditions on φ cannot be replaced by δ-quasiconvexity.

(b) In Theorems 2.1 and 2.2, the conclusions may fail if φ is δ-convex but the δ-local conditions are replaced by δ'-local conditions with $\delta' < \delta$.

To obtain examples in support of (a) and (b), suppose that $\alpha < \beta < \gamma$. For (a), let φ be constant with value σ on the closed interval $[\alpha, \beta]$ and constant with value $\eta > \sigma$ on the half-open interval $]\beta, \gamma]$. Then φ is quasiconvex and each point of its domain $[\alpha, \gamma]$ provides a local minimum for φ. However, the points of $]\beta, \gamma]$ actually provide global maxima rather than global minima. For (b), note that if $\gamma - \alpha < \delta$ then every real-valued function on $[\alpha, \gamma]$ is (trivially) δ-convex. If, in addition, $\gamma - \beta = \delta' < \delta$, there are many functions on $[\alpha, \gamma]$ for which γ provides a δ'-local strict minimum but either there is no global minimum or the global minimum is attained only in $[\alpha, \beta]$.

For another example in support of (b), let $C = [-1, 2[$ and let the function φ on C be such that

$$\varphi(-1) = 1, \quad \varphi(t) = 0 \quad \text{for } -1 < t < 1, \quad \varphi(t) \geqq t \quad \text{for } 1 \leqq t < 2.$$

Then φ is strictly 2-convex, and for $0 < \delta' < 2$ it has a δ'-local minimum at the point $\delta' - 1$ but does not have a global minimum there. If φ is modified to give it the value $-\frac{1}{2}$ at the point $\delta' - 1$, then φ is still strictly 2-convex and now has a δ'-local strict minimum at $\delta' - 1$.

3. Qualitative properties of $\rho(C, c)$. For the results of § 2, the convex domain C need not be closed and it may lie in an arbitrary normed linear space. However, from now on its convenient to work with the following.

STANDING HYPOTHESES. *E is a finite-dimensional normed vector space and C is a body in E, meaning that C is closed, convex, has nonempty interior, and does not contain any line. The sets of extreme and nonextreme points of C are denoted, respectively, by C_e and C_m.*

It is known that C_e is nonempty, and in fact C is the convex hull of the union of its extreme points and extreme rays [10]. For each $c \in C$, let

$$\xi(C, c) = \inf\{\|c - p\| : p \in C_e\},$$

the distance from the point c to the set C_e. This is a measure of how close the point c comes to being an extreme point of C. Another such measure is given by $\mu(C, c)$, defined as zero when $c \in C_e$ and defined for $c \in C_m$ as half the length of a longest segment in C that has c as its midpoint. (The existence of such a segment follows from an easy argument using the compactness of E's unit sphere, the closedness of C, and the fact that C contains no line.) Note that $\mu(C, c)$ is the smallest number σ such that each line L in E through c includes an endpoint of $L \cap C$ at distance at most σ from c.

When C is fixed, the function $\xi(C, c)|_{c\in C}$ will be denoted by $\xi(C, \cdot)$; similarly for the functions $\mu(C, c)|_{c\in C}$ and $\rho(C, c)|_{c\in C_m}$, where

$$\rho(C, c) = \frac{\xi(C, c)}{\mu(C, c)}.$$

Of course, $\rho(C, \cdot) \equiv 1$ when C is one-dimensional, but in general the two measures of near-extremeness are different and the behavior of the function $\rho(C, \cdot)$ is of interest. As was explained in the Introduction, the extreme value

$$\rho(C) = \sup\{\rho(C, c): c \in C_m\}$$

plays an essential role in Theorem 4.3's extension of property (P2) to δ-convex functions. Hence the extreme value

$$\rho(E) = \sup\{\rho(C): \text{body } C \subset E\}$$

is also of interest, and it is estimated in § 5.

It would be of interest, under various assumptions concerning the body C, to know the complexity of computing $\xi(C, c)$, $\mu(C, c)$, $\rho(C, c)$, and $\rho(C)$ and also, when C is bounded, the complexity of computing

$$\xi(C) = \sup\{\xi(C, c): c \in C\} \quad \text{and} \quad \mu(C) = \sup\{\mu(C, c): c \in C\}.$$

Of course, $\xi(C) \leqq \frac{1}{2}\delta(C)$ and $\mu(C) \leqq \frac{1}{2}\delta(C)$, where $\delta(C)$ is C's diameter. When C is a polytope presented in terms of its vertices, $\delta(C)$ and $\xi(C, c)$ are easy to compute but computation of the other numbers appears to be difficult. For a polytope presented in terms of its bounding hyperplanes, the computation of $\delta(C)$ is NP-hard in l^p-spaces with $1 \leqq p < \infty$ [6], and the same may well be true of the other numbers.

The remainder of the present section is devoted to some qualitative properties of the functions $\xi(C, \cdot)$, $\mu(C, \cdot)$, and $\rho(C, \cdot)$. For these properties, it suffices to deal with the usual Euclidean norm for $\mathbb{R}^d$, since each norm for $\mathbb{R}^d$ is caught between two positive multiples of the Euclidean norm. However, the quantitative details in §§ 4 and 5 depend on the choice of norm.

THEOREM 3.1. *For each body C, the function $\xi(C, \cdot)$ is everywhere continuous and the function $\mu(C, \cdot)$ is everywhere upper semicontinuous.*

Proof. Routine use of the triangle inequality shows that $|\xi(C, x) - \xi(C, y)| \leqq \|x - y\|$, whence of course $\xi(C, \cdot)$ is continuous. The upper semicontinuity of $\mu(C, \cdot)$ follows from a simple argument based on the closedness of C and the compactness of the unit sphere in the containing space. □

For each $d \geqq 3$, there exists a compact body $C \subset \mathbb{R}^d$ whose set C_e of extreme points is not closed. (For example, let C be the convex hull of the union of a $(d-1)$-dimensional spherical ball B and a segment S that is orthogonal to B's hyperplane and is centered at a point p in the boundary A of B. Then $p \in C_m$ but $A \setminus \{p\} \subset C_e$.) Clearly the function $\mu(C, \cdot)$ is discontinuous at each point p of C_m that belongs to the closure of C_e, and in fact the restriction of $\mu(C, \cdot)$ to $C_m \cup \{p\}$ is also discontinuous at p.

Although the function $\mu(C, \cdot)$ need not be everywhere continuous, continuity at certain points can be established. The following lemma is useful for that purpose. (As the terms are used here, a *polyhedron* is a set that is the intersection of a finite number of closed halfspaces: a *polytope* is a bounded polyhedron.)

LEMMA 3.2. *Suppose that C is a body in $\mathbb{R}^d$, the origin 0 is the midpoint of an open segment $]-q, q[$ in C, and H is the hyperplane through 0 orthogonal to the segment. If*

C is a polyhedron or the origin is interior to C, then for each $\lambda \in]0, 1[$ there is a neighborhood U of 0 in $C \cap H$ such that the vector sum

$$U + \lambda\,]-q, q[= \{u + \lambda q : u \in U, |\tau| < \lambda\}$$

is a neighborhood of 0 in C.

Proof. The case in which $0 \in \operatorname{int} C$ is left to the reader. In the remaining case, there is a representation of C in the form $C = \bigcap_{i=1}^{n} K_i$, where each K_i is a closed halfspace with bounding hyperplane J_i and where, for some $m \geqq 1$, $0 \in \bigcap_{i=1}^{m} J_i$ and $0 \in \operatorname{int} \bigcap_{i=m+1}^{n} K_i$. By a well-known theorem [5], [11], the polyhedron $K = \bigcap_{i=1}^{n} K_i$ is the direct sum of the subspace $J = \bigcap_{i=1}^{m} J_i$ and the polyhedral cone $K \cap J^{\perp}$, where $J^{\perp}$ is the orthogonal complement of J. Of course, $[-q, q] \subset J$, whence $J^{\perp} \subset H$. Since the segment $[-\lambda q, \lambda q]$ is interior to the intersection $\bigcap_{i=m+1}^{n} K_i$, there is a positive η such that this intersection contains the vector sum of the segment $[-\lambda q, \lambda q]$ and the η-neighborhood V of the origin in $\mathbb{R}^d$. Then the set $U = V \cap C \cap H$ is the desired neighborhood of 0 in $C \cap H$. □

THEOREM 3.3. *The function $\mu(C, \cdot)$ is continuous at each interior point of C, and if C is polyhdral then $\mu(C, \cdot)$ is continuous everywhere.*

Proof. By Theorem 3.1, the function $\mu(C, \cdot)$ is everywhere upper semicontinuous. To complete the proof, we show that if p is interior to C, or $p \in C$ and C is polyhedral, then $\mu(C, \cdot)$ is lower semicontinuous at p. This is obvious when $\mu(C, p) = 0$ for the function μ is nonnegative. Suppose then that $\mu(C, p) > 0$, assume without loss of generality that $p = 0$, and let $[-q, q]$ be a longest segment in C that has 0 as midpoint. Now consider an arbitrary $\lambda \in]0, 1[$, and let U be as in Lemma 3.2. Then for each $\varepsilon \in]0, \lambda[$, the set $W = U + \varepsilon\,]-q, q[$ is a neighborhood of 0 in C, and for each $w \in W$ it is true that

$$w + (\lambda - \varepsilon)\,]-q, q[\subset U + \lambda\,]-q, q[\subset C$$

and hence $\mu(C, w) \geqq (\lambda - \varepsilon)\|q\|$. This shows that the function $\mu(c, \cdot)$ is lower semicontinuous at p. □

LEMMA 3.4. *If the body C is a pointed cone, then the function $\rho(C, \cdot)$ attains a positive minimum on C. If C is a pointed polyhedral cone, then $\rho(C, \cdot)$ also attains a maximum on C.*

Proof. We assume without loss of generality that the origin is the apex of C, whence there is a hyperplane H that misses the origin and there is a compact body B in H such that $C = [0, \infty[\,B$. Since C is a cone, $C_m = C \setminus \{0\}$, and for each $c \in C_m$ and $\lambda > 0$ it is true that $\xi(C, \lambda c) = \lambda \xi(C, c)$ and $\mu(C, \lambda c) = \lambda \mu(C, c)$. Hence the range of the function $\rho(C, \cdot)$ on C_m is equal to its range on B.

On the set B, the functions $\xi(C, \cdot)$ and $\mu(C, \cdot)$ are both positive. By Theorem 3.1, $\xi(C, \cdot)$ is continuous and $\mu(C, \cdot)$ is upper semicontinuous. Hence their quotient $\rho(C, \cdot)$, being lower semicontinuous and positive on the compact set B, attains a positive minimum on B.

If C is polyhedral then, by Theorem 3.3, the function $\mu(C, \cdot)$ is actually continuous on C, whence $\rho(C, \cdot)$ is continuous and hence attains a maximum on the compact set B. □

LEMMA 3.5. *If the body C is unbounded and its set C_e of extreme points is bounded, then*

$$\limsup_{n \to \infty} \{\rho(C, c) : c \in C_m, \|c\| \geqq n\} = 1.$$

Proof. We may assume without loss of generality that the origin 0 belongs to the bounded set C_e. Let K denote the union of all rays that issue from the origin 0 and are

contained in C, and let B denote the closed convex hull of C_e. Then K is a pointed closed convex cone, B is a compact convex set, and $C = B + K$. Let $\beta = \sup\{\|b\|: b \in B\}$. Since $0 \in C_e$, it is true for each point $c \in C$ that $\xi(C, c) \leqq \|c\|$. Each point $c \in C \setminus B$ has at least one representation in the form $c = b + k$ with $b \in B$ and $k \in K \setminus \{0\}$. For each such representation, it is true that $b + 2k \in C$, whence $c \in C_m$ with $\mu(C, c) \geqq \|k\|$ and hence

$$\rho(C, c) \leqq (\|k\| + \beta)/\|k\|.$$

From this it follows that

$$\limsup_{n\to\infty} \{\rho(C, c): c \in C_m, \|c\| \geqq n\} \leqq 1.$$

When the point $c \in K \setminus B$ belongs to R, $[0, 2c]$ is the unique longest segment in C that has c as its midpoint, and hence

$$\rho(C, c) \geqq (\|k\| - \beta)/\|c\| = 1.$$

This implies that

$$\limsup_{n\to\infty} \{\rho(C, c): c \in C_m, \|c\| \geqq n\} \geqq 1. \qquad \square$$

THEOREM 3.6. *If the body C is a polyhedron, then the function $\rho(C, \cdot)$ attains a maximum $\rho(C) \geqq 1$.*

Proof. Consider an arbitrary edge or extreme ray R of C, and an endpoint p of R. Then $p \in C_e$ and the set C_e is finite, so for each point $c \in R \setminus \{p\}$ sufficiently close to p it is true that $c \in C_m$ and

$$\xi(C, c) = \|c - p\| = \mu(C, c).$$

This implies that the function $\rho(C, \cdot)$ attains the value 1.

Now suppose that the function $\rho(C, \cdot)$ does not attain a maximum on c_m, and let $c_1, c_2, \cdots$ be a sequence in C_m such that

$$\rho(C, c_n) \to \rho(C) \quad \text{as } n \to \infty.$$

If $\|c_n\| \to \infty$, then $\rho(C) \leqq 1$ by Lemma 3.5, and the desired conclusion follows from the preceding paragraph.

In the remaining case, we may assume that the sequence $c_1, c_2, \cdots$, converges to a point $p \in C$. If $p \in C_m$ then $\rho(C, c) = \rho(C)$ by Theorem 3.3. If $p \in C_m$, then $p \in C_e$. It is not hard to verify that the values of $\rho(C, c)$ attained when the point $c \in C_m$ belongs to a sufficiently small neighborhood of the extreme point p are precisely the values attained by the function $\rho(K, \cdot)$ where K is the cone consisting of all rays that issue from p and pass through the various points of C. When C is a polyhedron, this cone K is also polyhedral and the function $\rho(K, \cdot)$ attains a maximum by Lemma 3.4. $\square$

To end this section, we note that for each two-dimensional body C, the following four conditions are equivalent: (a) C is a polyhedron; (b) the set C_e is finite; (c) the function $\rho(C, \cdot)$ attains a minimum; (d) the infimum of $\rho(C, \cdot)$ is positive. Of course, (a) implies (b)-(d) in any dimension. However, for bodies in $\mathbb{R}^3$, (b) & (c) & (d) does not imply (a), (c) does not imply (b) or (d), and some other questions about implications among these conditions are unsettled. In particular, we do not know whether (d) implies (c) nor whether (d) implies (b).

4. Maxima. The following remark illustrates the manner in which the quantity $\rho(C)$ enters into the δ-version of (P2).

LEMMA 4.1. *If φ is a midpoint strictly δ-quasiconvex function on a body C, and q is a point of C at which φ attains its maximum, then there is an extreme point x of C such that $\|q-x\| \leqq \delta\rho(C)$.*

Proof. It suffices to show that $\mu(C, q) \leqq \delta$, for then

$$\xi(C, q) = \mu(C, q)\rho(C, q) < \delta\rho(C, q).$$

In the contrary case, q is the midpoint of a segment $[p, r]$ in C of length exceeding δ. But then, by the definition of midpoint strict δ-quasiconvexity, $\varphi(p) < \varphi(q)$ or $\varphi(r) > \varphi(q)$. This contradicts the assumed maximizing property of q. □

Our main tool for dealing with maxima is the following lemma from [12].

LEMMA 4.2. *Suppose that K is a convex set and the partial ordering $\prec$ on K is defined as follows: $v \prec w$ if and only if, for the line L through v and w, it is true that v is an inner point and w is an endpoint of the intersection $L \cap K$. (Intuitively, w can see beyond v in K, but v cannot see beyond w.) With respect to this ordering, each linearly ordered subset of K is affinely independent.*

This lemma yields a quick proof of the theorem of Hirsch and Hoffman [8] (essentially (P2) of the Introduction) asserting that if a convex function attains its maximum on a finite-dimensional line-free closed convex set, then it does so at an extreme point. As we now show, it also yields extensions of this result to functions that are globally convex or quasiconvex.

THEOREM 4.3. *Suppose that φ is a real-valued function defined on the body C. Suppose also that φ is midpoint δ-quasiconvex and C is compact, or that φ is midpoint δ-convex and is bounded above on each ray in C. Then for each value α of φ there are points q and x of C such that $\varphi(q) \geqq \alpha$, x is an extreme point of C, and $\|q-x\| < \delta\rho(C)$.*

Proof. In terms of the ordering $\prec$ described in Lemma 4.2, we define a second ordering $\ll$ by saying that $v \ll w$ provided that $v \prec w$ and $\varphi(v) \leqq \varphi(w)$. Every set that is linearly ordered by $\ll$ is also linearly ordered by $\prec$, and hence by Lemma 4.2 is of cardinality at most $d+1$ where d is the dimension of C. Now consider an arbitrary point $c \in C$ such that $\varphi(c) = \alpha$, and let

$$c = c_0 \ll c_1 \ll \cdots \ll c_k \tag{14}$$

be a $\ll$-ordered sequence that starts with the point c and cannot be extended. Set $q = c_k$, whence of course $\varphi(q) \geqq \alpha$. To complete the proof it suffices to show that $\mu(C, q) < \delta$, for then $q \in C_e$ or we have

$$\xi(C, q) = \frac{\xi(C, q)}{\mu(C, q)}\mu(C, q) = \rho(C, q)\mu(C, q) < \rho(C)\delta.$$

Suppose that $\mu(C, q) \geqq \delta$, let p and r be the endpoints of a longest segment in C that has q as a midpoint, and let L denote the line through p and r. At least one of p and r must be an endpoint of the intersection $L \cap C$, and we may assume that p is such an endpoint. If the intersection $L \cap C$ is a segment $[p, s]$ (where of course $r \in [p, s]$), it follows from the midpoint δ-quasiconvexity of φ that $\varphi(p) \geqq \varphi(q)$ or $\varphi(s) \geqq \varphi(q)$. Then the sequence (14) can be extended by adding p or s at the end, and the assumed maximality of the sequence is contradicted.

In the remaining case, the intersection $L \cap C$ is a ray R that issues from p and passes through r. This ray includes the point $r_k = q + k(q-p)$ for each integer $k \geqq -1$. From the midpoint δ-convexity of φ it follows that for each $k \geqq 1$,

$$\varphi(r_k) \geqq \varphi(r_{k-1}) + (\varphi(r_{k-1}) - \varphi(r_{k-2})),$$

and k successive applications of this inequality lead to the conclusion that

$$\varphi(r_k) \geqq (k+1)\varphi(r_0) - k\varphi(r_{-1}) = \varphi(q) + k(\varphi(q) - \varphi(p)).$$

Since, by hypothesis, φ is bounded above on the ray R, it follows that $\varphi(p) \geqq \varphi(q)$. But then the sequence (14) can be extended by adding p, and the contradiction completes the proof. □

Now suppose that the function $\rho(C, \cdot)$ attains a maximum on C at a point $c \in C_m$. (By Theorem 3.6, this occurs whenever C is a polytope.) Let $\delta = \mu(C, c)$, whence of course $\xi(C, c) = \delta\rho(C)$. Then Theorem 4.3 is sharp for C in the following sense.

For each $\eta > \delta$ there exists a continuous η-convex function φ on C such that the maximum of φ is attained only at c. Hence δ is the largest "modulus of global convexity" ε such that the ε-convexity of a continuous φ ensures the existence of an extreme point of C within $\delta\rho(C)$ of c.

To construct the desired function φ, start from an arbitrary continuous convex function ψ that is bounded above on C. Then let $\varphi = \psi + \zeta$, where

$$\zeta(x) = 1 - \frac{\|x - c\|}{\eta - \delta} \quad \text{for } \|x - c\| < \eta - \delta,$$

and $\zeta = 0$ elsewhere.

In the Introduction, we have mentioned a possible smoothing condition (7). A weakened form of this condition appears as a hypothesis in the following theorem.

THEOREM 4.4. *Suppose that the body C is compact and the function φ on C satisfies the following condition: each point $y \in C_m$ is an inner point of a segment $[x, z]$ in C such that $\varphi(y) \leqq \max\{\varphi(x), \varphi(z)\}$. Then for each value α of φ there is an extreme point x of C such that $\varphi(x) \geqq \alpha$. In particular, if φ attains a maximum on C then it does so at an extreme point.*

Proof. The proof follows the first two paragraphs of the proof of Theorem 4.3. □

5. Estimation of $\rho(E)$.

LEMMA 5.1. *For each $c \in C_m$ there exists a ball B centered at c such that $\rho(C, c) = \rho(C \cap B, c)$.*

Proof. Let x be an extreme point of C, and let $[p, r]$ have maximum length among the segments that are centered at c and contained in C. Let B be any ball that is centered at c and contains x, p, and r. Then of course $\mu(C \cap B, c) = \mu(C, c)$. Also, $\xi(C \cap B, c) = \xi(C, c)$, because each point of $(C \cap B)_e \setminus C_e$ belongs to the boundary of B and hence is at distance at least $\|x - c\|$ from c. □

LEMMA 5.2. *If C is bounded and $c \in C_m$, there exists a d-simplex S such that $c \in S$ and the vertices of S are extreme points of C. For each such S, it is true that $c \in S_m$ and $\rho(C, c) \leqq \rho(S, c)$.*

Proof. Since $C = \operatorname{con} C_e$, the existence of S follows from Carathéodory's Theorem [2]. It is obvious that $c \in S_m$, $\xi(C, c) \leqq \xi(S, c)$, and $\mu(C, c) \geqq \mu(S, c)$. □

THEOREM 5.3. *For each d-dimensional E, $\rho(E) \leqq d$.*

Proof. In view of Lemmas 5.1 and 5.2, it suffices to show that if the origin 0 belongs to a d-simplex S with vertices $v_0, \cdots, v_d$, and if $\|v_i\| \geqq 1$ for all i, then S contains a segment of length at least $2/d$ centered at the origin.

Since $0 \in S$, there are numbers $\lambda_i \geqq 0$ such that

$$\sum_{i=0}^{d} \lambda_i = 1 \quad \text{and} \quad \sum_{i=0}^{d} \lambda_i v_i = 0.$$

Then $\lambda_i \geqq 1/(d+1)$ for at least one value of i, and we may assume without loss of generality that this is λ_0. From the fact that $-\lambda_0 v_0 = \sum_{i=1}^{d} \lambda_i v_i$, it follows that the point

$$p = -\frac{\lambda_0}{1 - \lambda_0} v_0$$

is a convex combination of $\{v_1, \cdots, v_d\}$ and hence $p \in S$. But then S contains the segment $[p, v_0]$, and hence contains the segment $[-p, p]$ centered at 0. Since $\lambda_0 \geqq 1/(d+1)$, it is true that

$$\|p\| \geqq \frac{\lambda_0}{1-\lambda_0}\|v_0\| \geqq \frac{1/(d+1)}{1-1/(d+1)} = \frac{1}{d}. \qquad \square$$

For each positive integer d, and for $1 \leqq p < \infty$, let $\mathbb{R}_p^d$ denote the space $\mathbb{R}^d$ with the norm of

$$x = (x_1, \cdots, x_d) \in \mathbb{R}^d$$

given by

$$\|x\| = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p} \qquad (1 \leqq p < \infty).$$

Let $\mathbb{R}_\infty^d$ denote the space $\mathbb{R}^d$ with the norm

$$\|x\| = \max \{|x_1|, \cdots, |x_d|\}.$$

THEOREM 5.4. *If E is the d-dimensional subspace of $\mathbb{R}_\infty^{d+1}$ consisting of all points for which the sum of the coordinates is zero, then $\rho(E) = d$.*

Proof. Let S denote the d-simplex in R_∞^{d+1} consisting of all points $x = (x_1, \cdots, x_{d+1})$ such that $\sum_{i=1}^{d+1} x_i = 1$ and $x_i \geqq 0$ for all i. Then $\rho(S) \leqq d$ by Theorem 5.3. To complete the proof of Theorem 5.4, we show that $\rho(S, c) \geqq d$ for the centroid c of S. Since the distance from c to each extreme point of S is $d/(d+1)$, we want to show that whenever p is a point of the space such that $c+p \in S$ and $c-p \in S$, then $\|p\| \leqq 1/(d+1)$. But that is obvious, for if some coordinate of p exceeds $1/(d+1)$ in absolute value, then $c+p$ or $c-p$ has a negative coordinate and hence does not belong to S. $\square$

In the d-dimensional space E of Theorem 5.4, the unit ball is a hyperplane section of a $(d+1)$-cube. When $d=2$, it is a regular hexagon. In the general case, it is of the form $T \cap -T$, where T is a d-simplex whose centroid is the origin. Theorem 5.4 establishes the sharpness of Theorem 5.3 in the sense that for each d there exists a d-dimensional E for which $\rho(E) = d$. However, it is also of interest to determine (or at least find sharp estimates for) $\rho(E)$ for the "standard" spaces $\mathbb{R}_p^d (1 \leqq p \leqq \infty)$. The following lemma is useful in dealing with the case $p = \infty$.

LEMMA 5.5. *If E is a subspace of F, then $\rho(E) \leqq \rho(F)$.*

Proof. It suffices to consider the case in which E is a hyperplane through the origin in F. Let $p \in F \setminus E$. For each $\eta < \rho(E)$, there exists a compact convex body C in E and a point $c \in C_m$ such that $\rho(C, c) > \eta$. For each $\lambda > 0$, let $C_\lambda = C + [-\lambda, \lambda]p$, a compact convex body in F. Since it is clear that

$$\lim_{\lambda \to 0} \rho(C_\lambda, c) = \rho(C, c),$$

the desired conclusion follows. $\square$

COROLLARY 5.6. *For each d,*

$$d - 1 \leqq \rho(\mathbb{R}_\infty^d) \leqq d.$$

Proof. To prove the corollary we use Theorems 5.3 and 5.4 and Lemma 5.5. $\square$

Before turning to the case of Euclidean d-space $\mathbb{R}_2^d$, we establish one more geometric lemma that applies to all spaces and is of some interest in itself.

LEMMA 5.7. *For each $\eta < \rho(E)$, there exists in E a d-simplex S such that the origin is interior to S, each vertex of S is at distance 1 from the origin, and $\eta < \rho(S, 0)$.*

Proof. By the definition of $\rho(E)$ in conjunction with Lemmas 5.1 and 5.2, there exists a d-simplex $T \subset E$ and a point $t \in T_m$ such that $\rho(T, p) > \eta$. By an easy continuity argument, $\rho(T, s) > \eta$ for each interior point s of T sufficiently close to t. Now with s fixed, let x be an extreme point of T closest to s. We may assume without loss of generality that $s = 0$ and $\|x\| = 1$. Then for each vertex v of T, let

$$\bar{v} = \frac{v}{\|v\|} \in T,$$

and let S be the simplex whose vertices are the $\bar{v}$'s. It is clear that the pair $(S, 0)$ has the stated properties. □

LEMMA 5.8. *If S is a regular Euclidean d-simplex and s is the centroid of S, then*

$$\rho(S, s) = \begin{cases} \sqrt{d} & \text{when } d \text{ is odd,} \\ \sqrt{d+1} & \text{when } d \text{ is even.} \end{cases}$$

Proof. With $u_0, \cdots, u_d$ denoting the standard unit basis vectors for $\mathbb{R}_2^{d+1}$, let

$$c = \frac{1}{d+1} \sum_{i=0}^{d} u_1 \quad \text{and} \quad S = \operatorname{con}\{u_0 - c, \cdots, u_d - c\}.$$

Then S is a regular Euclidean d-simplex with centroid $s = 0$, and

$$\xi(S, s) = \left(\frac{d}{d+1}\right)^{1/2}.$$

We want to compute $\mu(S, s)$, which is half of the maximum of the norm on the set $S \cap -S$.

The points x of $S \cap -S$ are characterized by the existence of nonnegative numbers $\lambda_0, \cdots, \lambda_d$ with sum 1 and nonnegative numbers $\eta_0, \cdots, \eta_d$ with sum 1 such that

$$\sum_{i=0}^{d} \lambda_i (u_i - c) = x = \sum_{i=0}^{d} \eta_i (c - u_i).$$

From this it follows that

$$\sum_{i=0}^{d} (\lambda_i + \eta_i) u_i = 2c$$

and by linear independence of the u_i's that

$$\lambda_0 + \eta_0 = \lambda_1 + \eta_1 = \cdots \lambda_d + \eta_d = \frac{2}{d+1}.$$

Note also that since the origin is the orthogonal projection of the point c onto the hyperplane aff S, maximizing $\|x\|$ over $x \in S \cap -S$ is equivalent to maximizing $\|x - c\|^2$ over $x \in S \cap -S$. And for $x = \sum_{i=0}^{d} \eta_i u_i$ as described,

$$\|x - c\|^2 = \sum_{i=0}^{d} \eta_i^2.$$

We claim that for any x that maximizes $\|x\|^2$ over $S \cap -S$, there is at most one index i for which the numbers λ_i and η_i are both positive. For suppose there are two such indices, say 1 and 2 with $\eta_1 \geqq \eta_2$. Then for a sufficiently small positive ε we may increase η_1 and λ_2 by ε and decrease η_2 and λ_1 by ε without violating the above conditions, and since

$$(\eta_1 + \varepsilon)^2 - (\eta_2 - \varepsilon)^2 - \sum_{i=0}^{d} \eta_i^2 = \sum_{i=0}^{d} \eta_i^2 + 2\delta(\eta_1 - \eta_2 + \varepsilon) > \sum_{i=0}^{d} \eta_i^2 = x - \|c\|^2,$$

the maximizing property of x is contradicted. It follows, therefore, that for each index i with at most one exception, either $\lambda_i = 0$ and $\eta_i = 2/(d+1)$ or $\lambda_i = 2/(d+1)$ and $\eta_i = 0$. And since $\sum_{i=0}^{d} \eta_i = 1$, no more than half of the $d+1$ η_i's can be equal to $2/(d+1)$.

From the above observations we see that when d is odd—say $d = 2n+1$—the maximum of $\|v\|$ for $v \in S \cap -S$ is attained by setting

$$v = \frac{1}{d+1}\left(\sum_{i=0}^{n} u_i\right) - c.$$

Then $\|v\| = 1/\sqrt{2n-2} = \sqrt{d+1}$ and $\rho(S, s) = \xi(S, s)/\|v\| = \sqrt{d}$.

When d is even—say $d = 2n$—there is a maximizing x such that for some $r \leqq n$,

$$\eta_i = \frac{2}{d+1} \quad \text{for } 0 \leqq i < r \quad \text{and} \quad \eta_i = 0 \quad \text{for } r < i \leqq d.$$

From the fact that

$$\frac{2r}{d+1} + \eta_r = 1 \quad \text{and} \quad 0 \leqq \eta_r \leqq \frac{2}{d+1},$$

it follows that $2r \geqq d-1$, whence $r = n$ and $\eta_r = 1/d$. Hence the point

$$w = \left(\frac{2}{d+1}\sum_{i=0}^{n-1} u_i + \frac{1}{d+1} u_n\right) - c$$

is a point of $S \cap -S$ farthest from the origin, and we have

$$\|w\| = \frac{\sqrt{2n}}{2n+1} = \frac{\sqrt{d}}{d+1} \quad \text{and} \quad \rho(S, s) = \frac{\xi(S, s)}{\|w\|} = \sqrt{d+1}. \qquad \square$$

For each dimension d, Lemma 5.8 provides a lower bound for $\rho(\mathbb{R}_2^d)$ and it may be that this bound is sharp. We are able to show that it is sharp for $d \leqq 2$ (i.e., $\rho(\mathbb{R}_2^2) = \sqrt{3}$), and for $d \geqq 3$ to establish an upper bound that is no more than $\sqrt{5}$ times the lower bound. It follows from Lemma 5.7 that in seeking an upper bound for $\rho(\mathbb{R}_2^d)$, we may confine our attention to the ratio $\rho(Z, 0)$, where Z satisfies the conditions of the following lemma. Under these conditions, $\xi(Z, 0) = 1$ and we seek a lower bound on $\mu(Z, 0)$.

LEMMA 5.9. *With $d \geqq 2$, suppose that the origin in $\mathbb{R}_2^d$ is interior to a d-simplex Z whose vertices $z_0, \cdots, z_d$ are all of unit norm, and that the positive numbers λ_i are such that*

$$\sum_{i=0}^{d} \lambda_i = 1 \quad \textit{and} \quad \sum_{i=0}^{d} \lambda_i z_i = 0.$$

Let

$$S = \operatorname{con}\{z_1, \cdots, z_d\} \quad \textit{and} \quad s = -\frac{\lambda_0}{1-\lambda_0} z_0 \in S.$$

Then the following statements are true:

(i) *S is the facet of Z opposite z_0;*

(ii) *s belongs to the interior of S relative to the hyperplane* aff S;

(iii) *for each i, $\lambda_i < \frac{1}{2}$;*

(iv) $S \cap -Z \supset (1-2\lambda_0)((S-s) \cap (s-S)) + s.$

(Thus the intersection $S \cap -Z$ contains a contraction in the ratio $1-2\lambda_0$ of the reflection of the set S in the point s.)

Proof. The assertion (i) is obvious, and (ii) follows from the fact that

$$s = -\frac{1}{1-\lambda_0}(\lambda_0 z_0) = -\frac{1}{1-\lambda_0}\left(-\sum_{i=1}^{d} \lambda_i z_i\right) = \sum_{i=1}^{d} \frac{\lambda_1}{1-\lambda_0} z_i.$$

For (iii), note that if $\langle \cdot \rangle$ denotes the usual inner product, then

$$\lambda_0 = \lambda_0 \langle z_0, z_0 \rangle = \langle z_0, \lambda_0 z_0 \rangle = \left\langle z_0, -\sum_{i=1}^{d} \lambda_i z_i \right\rangle = \sum_{i=1}^{d} \lambda_i (-\langle z_0, z_i \rangle)$$

$$\leqq \sum_{i=1}^{d} \lambda_1 |\langle z_0, z_i \rangle| \leqq \sum_{i=1}^{d} \lambda_i = 1 - \lambda_0.$$

From this it follows that $\lambda_0 \leqq \frac{1}{2}$, with $\lambda_0 = \frac{1}{2}$ only if each of the points $z_1, \cdots, z_d$ is equal to $-z_0$. Equality is excluded by the assumption that the origin is interior to Z.

Now let $T = S - s$. To establish (iv), we observe that

$$\begin{aligned} S \cap -Z &\supseteq (T+s) \cap ((1-2\lambda_0)(-T-s+z_0) - z_0) \\ &= (T+s) \cap ((1-2\lambda_0)(-T)+s) \\ &\supseteq T \cap ((1-2\lambda_0)(-T)) + s \\ &\supseteq ((1-2\lambda_0)T) \cap ((1-2\lambda_0)(-T)) + s \\ &\supseteq (1-2\lambda_0)(T \cap -T) + s \\ &= (1-2\lambda_0)(S-s) \cap (s-S) + s. \end{aligned}$$

The first $\supseteq$ or $=$ follows from the fact that $S = T + s$ and

$$\begin{aligned} (1-2\lambda_0)(-T-s+z_0) - z_0 &= (1-2\lambda_0)(-S) + (2\lambda_0)(-z_0) \\ &\subset \operatorname{con}((-S) \cup \{-z_0\}) = -Z. \end{aligned}$$

For the second, use the fact that $s = (-\lambda_0/(1-\lambda_0))z_0$. The third, fifth, and sixth are obvious. For the fourth, note that since $0 < 1-2\lambda_0 < 1$ by (iii), and since the set T is convex and includes the origin, it is true that $T \supset (1-2\lambda_0)T$.

(Actually,

$$S \cap -Z = (T+s) \cap ((1-2\lambda_0)(-T)+s),$$

but we do not use this fact.) □

THEOREM 5.10. $\rho(\mathbb{R}_2^2) = \sqrt{3}$.

Proof. Let the notation be as in Lemma 5.9. With p denoting the foot of the perpendicular from the origin to the line aff S through z_1 and z_2, let the z_i's and λ_i's be relabeled to obtain $\|p\| \geqq \frac{1}{2}$. (It is easy to see that this is possible.) With

$$\delta = (1 - \|p\|^2)^{1/2} - \|s-p\|,$$

the situation is as shown in Fig. 2, and it follows from (iv) of Lemma 5.9 that there exists $q \in S \cap -Z$ such that

$$\|q-p\| \geqq \|s-p\| + (1-2\lambda_0)\delta.$$

Let us define $y = \|s-p\|$ and $r = \|p\|$, whence $\|s\| = (y^2+r^2)^{1/2}$ and it follows that

$$\lambda_0 = \frac{(y^2+r^2)^{1/2}}{1+(y^2+r^2)^{1/2}}.$$

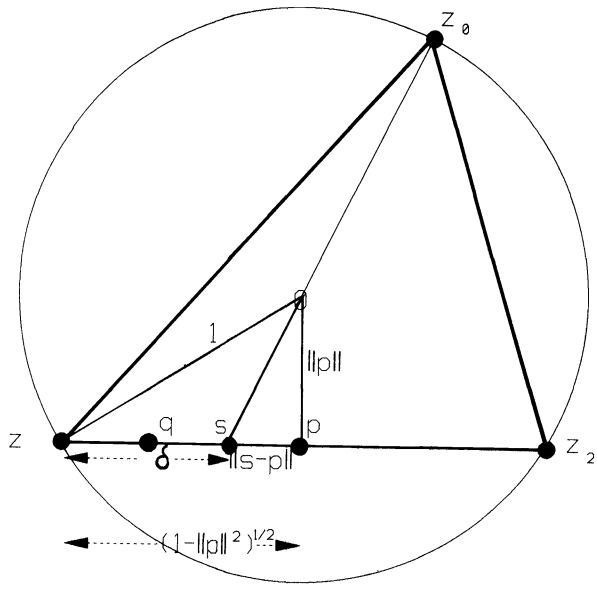

FIG. 2

Hence the lower bound on $\|q-p\|$ may be written as

$$g(y)=y+\left(1-\frac{2(y^2+r^2)^{1/2}}{1+(y^2+r^2)^{1/2}}\right)((1-r^2)^{1/2}-y)$$
$$=\frac{2(y^2+r^2)^{1/2}y}{1+(y^2+r^2)^{1/2}}+\left(1-\frac{2(y^2+r^2)^{1/2}y}{1+(y^2+r^2)^{1/2}}\right)(1-r^2)^{1/2}.$$

We claim that for each fixed $r\geqq\frac{1}{2}$, the value of $g(y)$ is minimized by setting $y=0$. Since

$$\frac{1}{2}(g(y)-g(0))=\frac{y(y^2+r^2)^{1/2}(1+r)+(r-(y^2+r^2)^{1/2})(1-r^2)^{1/2}}{(1+r)(1+(y^2+r^2)^{1/2})},$$

it will suffice to show that $f(y)\geqq f(0)$ for all $y\geqq 0$, where

$$f(y)=(y(1+r)-(1-r^2)^{1/2})(y^2+r^2)^{1/2}.$$

Now

$$f'(y)=(1+r)(y^2+r^2)^{1/2}+y(y^2+r^2)^{-1/2}(y(1+r)-(1-r^2)^{1/2})$$
$$=(2(1+r)y^2-y(1-r^2)^{1/2}+(1+r)r^2)(y^2+r^2)^{-1/2}$$

and since the discriminant $(1-r^2)-8(1+r)^2r^2$ is positive when $r\geqq\frac{1}{2}$ it follows that $f'(y)\geqq 0$ for all $y\geqq 0$. But then $f(y)\geqq 0$, and from this it follows that

$$g(y)\geqq g(0)=\frac{(1-r)^{3/2}}{(1+r)^{1/2}}.$$

Recalling the relevant definitions, we see that there is a point q of $S\cap -Z$ such that $\|q-p\|\geqq g(0)$, whence by the Pythagorean Theorem the squared norm of q is at least

$$h(r)=\frac{(1-r)^3}{1+r}+r^2.$$

Now

$$h'(r)=\frac{4}{(1+r)^2}(r^2+2r-1),$$

and this is positive when $r\geqq\frac{1}{2}$. Hence for $r\geqq\frac{1}{2}$,

$$h(r)\geqq\frac{(1/2)^3}{3/2}+\frac{1}{4}=\frac{1}{3},$$

whence the half-length of the segment $[-q, q]\subset Z\cap -Z$ is at least $1/\sqrt{3}$. This implies that $\rho(Z,0)\leqq\sqrt{3}$, whence $\rho(\mathbb{R}_2^2)\leqq\sqrt{3}$. The reverse inequality appears in Lemma 5.8. □

In preparation for the next result, we require a computational lemma.

LEMMA 5.11. *If the sequence* $\gamma_1, \gamma_2, \cdots$ *is defined by the condition that* $\gamma_1=1$ *and*

$$\gamma_d=\frac{1}{d^2}-\frac{(d-1)^4}{d^2(d+1)^2}\gamma_{d-1},$$

for $d\geqq 2$, *then*

$$\gamma_d=\frac{1}{15d^3(d+1)^2}(3d^4+15d^3-2d^2+15d+2)\geqq\frac{1}{5d}\left(1-\frac{1}{d}\right)^2.$$

Proof. Let $\alpha_d=d^2\gamma_d$, so that $\alpha_1=1$ and

$$\begin{aligned}\alpha_d&=1+\left(\frac{d-1}{d+1}\right)^2\alpha_{d-1}=1+\left(\frac{d-1}{d+1}\right)^2\left(1+\left(\frac{d-2}{d}\right)^2\alpha_{d-2}\right)\\&=1+\left(\frac{d-1}{d+1}\right)^2+\left(\frac{d-1}{d+1}\right)^2\left(\frac{d-2}{d}\right)^2\left(1+\left(\frac{d-3}{d-1}\right)^2\alpha_{d-3}\right)\\&=\frac{1}{d^2(d+1)}(2d^2(d^2+1)+(d-1)^2(d-2)^2+(d-2)^2(d-3)^2\alpha_{d-3})\\&=\frac{1}{d^2(d+1)^2}\left(2d^2(d^2+1)+\sum_{i=2}^{d-2}i^2(i+1)^2\right).\end{aligned}$$

A straightforward induction shows that

$$\sum_{i=1}^{n}i^2(i+1)^2=\frac{1}{15}n(n+1)(n+2)(3n^2+6n+1).$$

With $d=n-2$ this yields

$$\alpha_d=\frac{1}{15d(d+1)^2}(3d^4+15d^3+2d^2+15d+2)$$

and hence

$$\begin{aligned}\alpha_d=\frac{1}{d^2}\alpha&>\frac{1}{15d^3(d+1)^2}(3d^4-6d^2+3)=\frac{3}{15d^3(d+1)^2}(d+1)^2(d-1)^2\\&=\frac{1}{5d^3}(d-1)^2=\frac{1}{5d}\left(1-\frac{1}{d}\right)^2.\end{aligned}$$

□

THEOREM 5.12. *If* $d\geqq 2$, *then*

$$\frac{d}{d-1}\sqrt{5d}\geqq\rho(\mathbb{R}_2^d)\geqq\begin{cases}\sqrt{d} & \text{for odd } d,\\ \sqrt{d+1} & \text{for even } d.\end{cases}$$

Proof. The lower bounds $\sqrt{d}$ and $\sqrt{d+1}$ come from Lemma 5.8. To justify the upper bound, it is more convenient to work with $\beta_d = 1/\rho(\mathbb{R}_2^d)$. We already know that $\beta_1 = 1$ and $\beta_2 = 1/\sqrt{3}$. To prove

$$\beta_d \geqq \frac{d-1}{d\sqrt{5d}}$$

for a given $d \geqq 2$, it will suffice, in view of Lemma 5.11, to show that

$$\beta_k^2 \geqq \frac{1}{k^2} + \frac{(k-1)^4}{k^2(k+1)^2}\beta_{k-1}$$

for all $k \leqq d$. This will be accomplished by induction on d.

With the notation as in Lemma 5.9, let the z_i's and λ_i's be relabeled to obtain $\lambda_0 \geqq 1/(d+1)$, let p denote the orthogonal projection of the origin on the hyperplane aff S, and let

$$\delta = (1 - \|p\|^2)^{1/2} - \|s - p\|.$$

Of course,

$$\frac{\mu(S, s)}{\xi(S, s)} \geqq \beta_{d-1}$$

and with the aid of (iv) of Lemma 5.9 we see that

$$\mu(S \cap -Z, s) \geqq (1 - 2\lambda_0)\mu(S, s).$$

For $1 \leqq j \leqq d$ it follows from the Pythagorean Theorem that

$$\|p - v_j\|^2 + \|p\|^2 = \|v_j\|^2 = 1$$

and from the triangle inequality that

$$\|p - v_j\| \leqq \|p - s\| + \|s - v_j\|.$$

Hence $\xi(S, s) \geqq \delta$ and we have

$$\mu(S \cap -Z, s) \geqq (1 - 2\lambda_0)\mu(S, s) \geqq (1 - 2\lambda_0)\beta_{d-1}\delta.$$

That is, the set $S \cap -Z$ contains a segment that is centered at s and has half-length at least $(1 - 2\lambda_0)\beta_{d-1}\delta$. For at least one end of this segment, the squared distance from p is at least

$$\|p - s\|^2 + (1 - 2\lambda_0)^2\beta_{d-1}^2\delta^2$$

and from the origin it is at least

$$Q = \|p - 2\|^2 + (1 - 2\lambda_0)^2\beta_{d-1}^2\delta^2 + \|p\|^2.$$

To prove Theorem 5.12 it will suffice, in view of Lemmas 5.7 and 5.11 and the fact that $\xi(Z, 0) = 1$, to show that

$$Q \geqq \frac{1}{d^2} + \frac{(d-1)^4}{d^2(d+1)^2}\beta_{d-1}.$$

For notational convenience, we write

$$y = \|s\| = \frac{\lambda_0}{1 - \lambda_0} \geqq \frac{1}{d} \quad \text{and} \quad r = \|p\| \leqq y \leqq 1.$$

Then $\lambda_0 = y/(1+y)$, and Q becomes

$$h(y) = y^2 + \left(1 - \frac{2y}{1+y}\right)^2 \beta_{d-1}^2((1-r^2)^{1/2} - (y^2-r^2)^{1/2})^2$$
$$= y^2 + (1-y)^4 \beta_{d-1}^2((1-r^2)^{1/2} + (y^2-r^2)^{1/2})^{-2}.$$

For each fixed y, the value of $h(y)$ decreases with decreasing $|r|$. Hence $h(y) \geqq f(y)$, where

$$f(y) = y^2 + \frac{(1-y)^4}{(1+y)^2}\beta_{d-1}^2$$

and

$$f'(y) = (1+y)^{-3}(2y(1+y)^3 - 4(1-y)^3(1+y)\beta_{d-1}^2 - 2(1-y)^4\beta_{d-1}^2)$$
$$\geqq (1+y)^{-3}(2y(1+y)^3 - 6\beta_{d-1}^2) \quad \text{when } \frac{1}{d} \leqq y \leqq 1.$$

With $\beta_{d-1} \leqq 1/\sqrt{d-1}$ and $y \geqq 1/d$, we have $f'(y) \geqq 0$ and hence

$$f(y) \geqq \left(f\left(\frac{1}{d}\right)\right)^2 = \frac{1}{d^2} + \frac{(1-1/d)^4}{(1+1/d)^2}\beta_{d-1}. \qquad \square$$

For bodies C lying in certain Minkowski spaces E, Theorems 5.3, 5.4, 5.10, and 5.12 bound the number $\rho(C)$ from above. Because of the relevance of $\rho(C)$ to the maximization of δ-convex functions, it is also of interest to know how *small* $\rho(C)$ can be for bodies $C \subset E$. It is easy to see that whenever $\dim E \geqq 2$, there are pointed polyhedral cones $K \subset E$ that have arbitrarily small values for $\rho(K)$. (Simply let K have a "large opening.") However, the situation for bounded bodies is much less obvious, and hence the following observation seems worth including.

THEOREM 5.13. *For each $d \geqq 2$, the set $\{\rho(C)$: bounded body $C \subset \mathbb{R}_2^d\}$ is equal to the interval $]0, \rho(\mathbb{R}_2^d)]$. In particular, when $d = 2$ it is the interval $]0, \sqrt{3}]$.*

Proof. We first show that if $\varepsilon > 0$, and if C_ε is the lens in the xy-plane formed by intersecting the disks of radius $R = (1+\varepsilon^2)/(2\varepsilon)$ centered at the points $(-R+\varepsilon, 0)$ and $(R-\varepsilon, 0)$, then $\rho(C_\varepsilon) \leqq \varepsilon$. Consider an arbitrary point $p = (x, y) \in C_\varepsilon$ with $x, y \geqq 0$. The point q of ∂C_ε nearest to p is the radial extension to ∂C_ε of p from the point $(-R+\varepsilon, 0)$. At q, ∂C has a tangent line L, and we consider the cap D that contains the point q and is cut from C_ε by the line M parallel to L and passing through the point $z = (0, 1)$. Let r denote the point at which M intersects the segment joining the point $(-R+\varepsilon, 0)$ to q. We consider separately the two possibilities: $p \in [r, q]$; $r \in [p, q]$.

Suppose first that $p \in [r, q]$, and let $\eta = \|p - q\|$. Then there is a chord of C_ε that passes through p, is centered at p, and is of length $2\sqrt{2\eta R - \eta^2}$. Hence

$$\rho(C_\varepsilon, p) \leqq \frac{\eta}{\sqrt{2\eta R - \eta^2}} \leqq \varepsilon.$$

Now suppose that $r \in [p, q]$, and note that if the lens subtends an angle θ at z, then $\tan\theta = 1/(-\varepsilon)$. Consequently, if η is as before, then $\eta = \psi + \phi$ where $\psi = \|p - r\|$, $\phi = \|r - q\|$, and

$$\|p - r\| = \psi \leqq \|z - r\|(\tan\theta) = (R\phi - \phi^2)^{1/2}/(R - \phi).$$

Now C_ε contains the chord $[z, p - z]$ centered at p, and the square of half the length of this chord is given by

$$\|p - r\|^2 + \|z - r\|^2 \geqq \|z - r\|^2 = 2R\phi - \phi^2.$$

As $\eta = \psi + \phi$, we have

$$\eta \leqq \phi + (R\phi - \phi^2)^{1/2}/(R - \phi).$$

For small ε,

$$\phi \geqq (R\phi - \phi^2)^{1/2}/(R - \phi)$$

and consequently $\phi \geqq \eta/2$. Thus C_δ contains a chord centered at p of half-length at least $R\eta - \frac{1}{4}\eta^2$, and it follows that

$$\rho(C_\varepsilon) \leqq \eta(R\eta - \tfrac{1}{4}\eta^2)^{1/2} \leqq \varepsilon.$$

That completes the discussion of the case $d = 2$.

Now for $d \geqq 3$, write $\mathbb{R}_2^d = \mathbb{R}_2^2 \times \mathbb{R}_2^{d-2}$ in the usual way, so that each point $p \in \mathbb{R}_2^d$ can be written as $\mathrm{p} = (\mathrm{p}', \mathrm{p}'')$ with $p' \in \mathbb{R}_2^2$ and $p'' \in \mathbb{R}_2^{d-2}$. Let μ be the gauge-function of the body $C_\varepsilon \subset \mathbb{R}^2$, let $\| \cdot \|$ denote the Euclidean norm for $\mathbb{R}^{d-2}$, and define

$$K_\varepsilon = \{p: (\mu(p'))^2 + \|p''\|^2 \leqq 1\}.$$

Then K_ε is a body in $\mathbb{R}^d$, and it is not hard to verify that $\rho(K_\varepsilon) \leqq \varepsilon$. (When $d = 3$, K_ε is obtained by rotating the set C_ε about its axis of symmetry.)

To complete the proof, note that by a simple continuity argument, it is true for each space E that the set

$$\{\rho(C): \text{bounded body } C \subset E\}$$

is a connected subset of $\mathbb{R}$. □

REFERENCES

[1] M. AVRIEL, W. E. DIEWERT, S. SCHAIBLE, AND W. ZIEMBA, *Introduction to concave and generalized concave functions*, in Generalized Convexity and Optimization in Economics, Academic Press, New York, 1981, pp. 21-50.

[2] C. CARATHÉODORY, *Über den Variabilitätsbereich des Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen*, Math. Ann., 64 (1907), pp. 95-115.

[3] W. E. DIEWERT, M. AVRIEL, AND I. ZANG, *Nine kinds of quasiconcavity and concavity*, J. Econom. Theory, 25 (1981), pp. 397-420.

[4] W. GINSBERG, *Concavity and quasiconcavity in economics*, J. Econom. Theory, 6 (1973), pp. 596-605.

[5] A. J. GOLDMAN, *Resolution and separation theorems for polyhedral convex sets*, Ann. of Math. Stud., 38 (1956), pp. 41-51.

[6] P. GRITZMANN AND V. KLEE, *Computational complexity of inner and outer j-radii of polytopes in finite-dimensional normed spaces*, in preparation.

[7] W. M. HIRSCH AND G. B. DANTZIG, *The fixed charge problem*, Naval Res. Logist. Quart., 15 (1968), pp. 413-424.

[8] W. M. HIRSCH AND A. J. HOFFMAN, *Extreme varieties, concave functions, and the fixed charge problem*, Comm. Pure Appl. Math., 14 (1961), pp. 355-369.

[9] S. KIRKPATRICK, C. D. GELATT, JR., AND M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671-680.

[10] V. KLEE, *Extremal structure of convex sets*, Arch. Math., 8 (1957), pp. 234-240.

[11] ———, *Some characterizations of convex polyhedra*, Acta Math., 102 (1959), pp. 79-107.

[12] ———, *Extreme points of convex sets without completeness of the scalar field*, Mathematika, 10 (1964), pp. 59-63.

[13] P. M. PARDALOS AND J. B. ROSEN, *Constrained Global Optimization*, G. Goos and J. Hartmanis, eds., Lecture Notes in Computer Science, 268, Springer-Verlag, Berlin, New York, 1987.

[14] J. PONSTEIN, *Seven kinds of convexity*, SIAM Rev., 9 (1967), pp. 115-119.

[15] S. SCHAIBLE AND W. ZIEMBA, EDS., *Generalized Convexity and Optimization in Economics*, Proc. NATO Advanced Study Institute, University of British Columbia, Vancouver, British Columbia, Canada, August 4-15, 1980, Academic Press, New York, 1981.

[16] P. J. M. VAN LAARHOVEN AND E. H. L. AARTS, *Simulated Annealing: Theory and Applications*, D. Reidel, Dordrecht, Boston, 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1048-1071, September 1989

009

APPLICATIONS OF OPTIMAL MULTIPROCESSES*

FRANK H. CLARKE† AND RICHARD B. VINTER‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. In a previous paper a theory of necessary conditions for optimal multiprocess problems was supplied, that is, problems in dynamic optimization involving a collection of control systems coupled through constraints on the endpoints of the trajectories and the cost functions. Here the scope of the new theory is illustrated by application to optimization problems in investment planning, impulse control, optics, robotics, and renewable resources.

Key words. optimal control, necessary conditions, differential inclusions

AMS(MOS) subject classifications. 49B10, 49B99

1. Introduction. Optimal multiprocess problems are dynamic optimization problems involving a collection of control systems, coupled through constraints in the endpoints of the constituent state trajectories and through the cost function. Such problems arise, for example, in flight mechanics (optimal control of multistage rockets [11]), optimal investment [21], routing problems over variable domains [13], [16], and plasma control [22]. To date, necessary conditions of optimality and minimization algorithms for these problems have been studied, for the most part, by ad hoc techniques on a case-by-case basis.

Our earlier paper provided a unified theory of necessary conditions for optimal multiprocess problems. The present paper is a companion piece in which we explore applications of the theory. On the one hand, we deduce from the general theory necessary conditions of optimality for certain special classes of problems of interest. On the other hand, we use the theory to study specific problems in detail and to determine the optimal multiprocesses involved.

In selecting problems for inclusion in this paper, our object has not been exclusively to give instances where established methods fail and we must have recourse to the theory of optimal multiprocesses. Indeed both the free time investment problem of § 4 and the robot arm problem of § 7 can be solved by traditional "two stage" techniques in which we freeze certain choice parameters, thereby reducing the problems to ones that can be solved by application of the classical Pontryagin maximum principle; we then minimize a value function (the minimum cost as a function of the frozen parameter values) to determine the optimum parameter values. The emphasis is rather on *interpretation* of conditions in the optimal multiprocess maximum principle. A recurring theme is that the "extra" conditions in our maximum principle correspond to stationarity of the value function.

Necessary conditions in optimization, in addition to providing the basis for solution of certain problems by analytical means, serve as inspiration for optimization algorithms and give information concerning asymptotic properties of sequences generated by these algorithms. These aspects of necessary conditions are illustrated, in the context of optimal control theory, in, for example, [15] and [25]. We expect that the optimal multiprocess maximum principle will also be the source of optimization algorithms,

* Received by the editors November 16, 1987; accepted for publication (in revised form) August 23, 1988.
† Centre de Recherches Mathématiques, Université de Montréal, Montréal, Quebec, Canada H3C 3J7.
‡ Department of Electrical Engineering, Imperial College, London SW7 2BT, United Kingdom.

in the case of optimal multiprocess problems. Now outside the simplest cases we cannot reasonably expect convenient formulae for the appropriate value functions to be available to us. It is significant therefore that the optimal multiprocess maximum principle makes no explicit reference to value functions, and implementation of optimization procedures associated with it will not depend on availability of such formulae in any way.

An important feature of our theory is that it treats dynamic optimization problems where the data is merely measurable in the time variable. Our purpose in treating the investment problem of § 4 is in part to illustrate a free time problem where we can expect the optimal free time to coincide with a point of discontinuity of the data, and that the optimal multiprocess maximum principle can provide useful information in such instances. When the data is discontinuous in the time variable, the standard boundary condition on the maximized Hamiltonian function is replaced by an inclusion involving the "essential values" of this function. On the face of it the essential value condition is rather weak, and it is somewhat surprising that it can supply quite precise information about solutions to dynamic optimization problems. It does so in the problem in § 4, for example. Drawing an analogy with convex optimization is helpful here. The condition "$0 \in \partial f(x)$," where ∂ denotes the subdifferential, appears a meager condition for x to be a minimizer of the convex function $f: \mathbb{R}^n \to \mathbb{R}$. But on the contrary it is necessary and sufficient. The point is that if the convex function f has a discontinuous derivative at a point x, then x is a rather likely location for a minimizer, and this is reflected in $\partial f(x)$ being set-valued and the optimality condition being somewhat unrestrictive. Likewise for free time dynamic optimization problems, a time at which the cost integral and/or dynamics are discontinuous is a likely location for the optimal free time, and an optimality condition in the form of an inclusion is not inappropriate.

The form this paper takes is such as to make it a convenient reference in further applications of the theory of optimal multiprocesses. It includes a number of useful results for computing the normal cones typically encountered. The optimal multiprocess maximum principle of [6] has been recast to include the kind of integral cost terms that often arise, and a slightly weaker version of the transversality condition than that in [6] is given, but one which is more convenient for many applications.

2. Preliminaries. Frequent reference is made to generalized gradients and normal cones. These are understood in the sense of Clarke [4].

DEFINITION 2.1. Let N be an open subset of $\mathbb{R}^k$, let x be a point in N, and let $f: N \to \mathbb{R}$ be a locally Lipschitz continuous function. Then the generalized gradient $\partial f(x)$ of f at x is the set

$$\partial f(x) = \overline{\text{co}} \left\{ \lim_i \nabla f(x_i) \,|\, x_i \to x, \quad \nabla f(x_i) \text{ exists for } i = 1, 2, \cdots \right\}.$$

Given a closed set $S \subset \mathbb{R}^k$, $d_C: \mathbb{R}^k \to \mathbb{R}$ denotes the Euclidean distance function

$$d_C(x) = \min_{y \in C} |y - x|.$$

DEFINITION 2.2. Let $C \subset \mathbb{R}^k$ be a closed subset of $\mathbb{R}^k$ and x a point in C. Then the normal cone to C at x, written $N_C(x)$, is

$$N_C(x) = \text{cl} \left\{ \bigcup_{\lambda \geqq 0} \lambda \, \partial d_C(x) \right\}. \tag{2.1}$$

Application of the theory of optimal multiprocesses usually involves analysis of

generalized gradients and normal cones. The following identities and estimates will be useful in this regard.

PROPOSITION 2.1. (i) *For any closed subset* $C \subset \mathbb{R}^k \times \mathbb{R}^l$ *and point* $(a, b) \in C$, *we have*

$$N_{\{(x,x,y)|(x,y)\in C\}}((a, a, b)) \subset \{(p, q, r) | (p+q, r) \in N_C((a, b))\}.$$

(ii) *For any closed subset* $C \subset \mathbb{R}^k \times \mathbb{R}^l$ *and point* (a, b, c) *such that* $(a-b, c) \in C$, *we have*

$$N_{\{(x,y,z)|(x-y,z)\in C\}}((a, b, c)) \subset \{(p, -p, r) | (p, r) \in N_C((a, b))\}.$$

(iii) *For closed sets* $C_1 \subset \mathbb{R}^k$ *and* $C_2 \subset \mathbb{R}^l$, *and points* $x \in C_1$ *and* $y \in C_2$ *we have*

$$N_{C_1 \times C_2}(x, y) = N_{C_1}(x) \times N_{C_2}(y).$$

(iv) *Let* $\tilde{f}: \mathbb{R}^k \times \mathbb{R}^l$ *be a given locally Lipschitz continuous function and take a point* $(a, b, c) \in \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^l$. *Define* $f: \mathbb{R}^k \times \mathbb{R}^k \times \mathbb{R}^l \to \mathbb{R}$ *to be*

$$f(x, y, z) = \tilde{f}(x-y, z).$$

Then

$$\partial f((a, b, c)) = \{(p, -p, q) | (p, q) \in \partial \tilde{f}((a-b, c))\}.$$

Properties (i) and (ii) follow from Theorem 2.5.6 of [4]. For (iii) see page 54 of [4]; Property (iv) is a consequence of the chain rule [4, Thm. 2.3.10].

We also call upon the concept of "essential values," introduced in [6].

DEFINITION 2.3. Let $I \subset \mathbb{R}$ be an open interval and let $g: I \to \mathbb{R}^k$ be a measurable function. Take a point $t \in I$. Then the set of essential values of g at t, written $\operatorname{ess}_{s\to t} g(s)$, comprises the points $x \in \mathbb{R}^k$ such that, for any $\varepsilon > 0$, the set

$$\{s | |x(s) - x| \leqq \varepsilon, |s - t| \leqq \varepsilon\}$$

has positive measure.

We remark that if the function g has finite left and right limits at t (written $g(t^-)$ and $g(t^+)$), then

$$\operatorname*{ess}_{s\to t} g(s) = \{g(t^-), g(t^+)\}$$

and so, in particular, if g is continuous, then

$$\operatorname*{ess}_{s\to t} g(s) = \{g(t)\}.$$

3. A maximum principle for optimal multiprocesses. The following data is given:

positive integers k and $n_i, m_i, \quad i = 1, \cdots, k,$

functions $\phi_i: \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{m_i} \to \mathbb{R}^{n_i},$

subsets U^i of $\mathbb{R} \times \mathbb{R}^{m_i}, \quad i = 1, \cdots, k,$

and subsets X^i of $\mathbb{R} \times \mathbb{R}^{n_i}, \quad i = 1, \cdots, k.$

We recall a notational device from [6]. A point $((a_1, b_1, \cdots), (a_2, b_2, \cdots), \cdots, (a_k, b_k, \cdots))$ is written $\{a_i, b_i, \cdots\}$.

A multiprocess is a k-tuple $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ each element in which comprises left and right endpoints τ_0^i, τ_1^i of a closed interval, an absolutely continuous function $y_i(\cdot):[\tau_0^i, \tau_1^i] \to \mathbb{R}^{n_i}$, and a measurable function $w_i(\cdot):[\tau_0^i, \tau_1^i] \to \mathbb{R}^{n_i}$ such that

$$\dot{y}_i(t) = \phi_i(t, y_i(t), w_i(t)), \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i],$$

$$w_i(t) \in U_t^i \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i],$$

$$y_i(t) \in X_t^i \quad \text{for all } t \in [\tau_0^i, \tau_1^i].$$

(Here X_t^i denotes $\{x: (t, x) \in X^i\}$, etc.)

The individual elements are called component processes, the $y_i(\cdot)$'s component trajectories, the $w_i(\cdot)$'s component control functions, and the intervals $[\tau_0^i, \tau_1^i]$ component time intervals.

Now let

$$L_i: \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{m_i} \to \mathbb{R}, \qquad i = 1, \cdots, k,$$

$$f: E \to \mathbb{R}$$

be given functions, in which $E = \prod_i (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{n_i})$, and let

$$\Lambda \subset \{\{\tau_0^i, \tau_1^i, y_0^i, y_1^i\} \in E \mid \tau_1^i \geqq \tau_0^i, i = 1, \cdots, k\}$$

be a given closed set. The optimal multiprocess problem is:

(P)
$$\text{minimize} \quad f(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\}) + \sum_i \int_{\tau_0^i}^{\tau_1^i} L_i(t, y_i(t), w_i(t))\, dt$$

over multiprocesses $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ satisfying $\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\} \in \Lambda$.

Define $\tilde{\phi}_i = [\phi_i, L_i]$, $i = 1, \cdots, k$. We invoke the following hypotheses.

(H1) For each $x \in \mathbb{R}^{n_i}$, $\tilde{\phi}_i(\cdot, x, \cdot)$ is $(\mathscr{L} \times \mathscr{B})$-measurable for $i = 1, \cdots, k$. Here $\mathscr{L}$ denotes the Lebesgue subsets in $\mathbb{R}$ and $\mathscr{B}$, the Borel subsets in $\mathbb{R}^{m_i}$.

(H2) U^i is a Borel measurable set for $i = 1, \cdots, k$.

(H3) There exists a constant K such that, for $i = 1, \cdots, k$, $|\tilde{\phi}_i(t, y, w)| \leqq K$ whenever $(t, y, w) \in \mathbb{R} \times X_t^i \times U_t^i$ and

(H4) $$|\tilde{\phi}_i(t, y, w) - \tilde{\phi}_i(t, y', w)| \leqq K|y - y'|$$
whenever $(t, y, w), (t, y', w) \in \mathbb{R} \times X_t^i \times U_t^i$.

(H5) f is locally Lipschitz continuous.

We now state a version of the optimal multiprocess maximum principle particularly suited to the applications to follow. Here, the functions H_i, $i = 1, \cdots, k$ are the Hamiltonian functions

$$H_i(t, x, u, p, \lambda) := p \cdot \phi_i(t, x, u) - \lambda L_i(t, x, u).$$

THEOREM 3.1. *Let $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ be an optimal multiprocess. Suppose that*

$$\text{graph}\,\{x_i(\cdot)\} \subset \text{interior}\,\{X^i\},$$

for $i = 1, \cdots, k$, and that hypotheses (H1)-(H5) *are satisfied. Then there exist a real number λ equal to zero or one, real numbers h_0^i, h_1^i, $i = 1, \cdots, k$, and absolutely continuous*

functions $p_i(\cdot):[T_0^i, T_1^i]\to\mathbb{R}^{n_i}$, $i=1,\cdots,k$, *such that* $\lambda+\sum_i|p_i(T_1^i)|>0$,

$$-\dot{p}_i(t)\in\partial_x H_i(t, x_i(t), u_i(t), p_i(t), \lambda), \qquad \text{a.e. } t\in[T_0^i, T_1^i],$$

$$H_i(t, x_i(t), u_i(t), p_i(t), \lambda)=\max_{w\in U_t^i} H_i(t, x_i(t), w, p_i(t), \lambda), \quad \text{a.e. } t\in[T_0^i, T_1^i],$$

$$h_0^i\in \operatorname{co}\operatorname*{ess}_{t\to T_0^i}\left[\sup_{w\in U_t^i} H_i(t, x_i(T_0^i), w, p_i(T_0^i), \lambda)\right],$$

$$h_1^i\in \operatorname{co}\operatorname*{ess}_{t\to T_1^i}\left[\sup_{w\in U_t^i} H_i(t, x_i(T_1^i), w, p_i(T_1^i), \lambda)\right],$$

for $i=1,\cdots$ *and*

$$\{-h_0^i, h_1^i, p(T_0^i), -p(T_1^i)\}\in N_\Lambda+\lambda\,\partial f.$$

The normal cone N_Λ *and the generalized gradient* ∂f *are evaluated at the point* $\{T_0^i, T_1^i, x(T_0^i), x(T_1^i)\}$.

Theorem 3.2 of [6] addresses optimal multiprocess problems with no integral cost terms and it incorporates a transversality condition expressed in terms of the generalized gradient of the distance function from Λ, ∂d_Λ. The Maximum Principle, Theorem 3.1, is derived from Theorem 3.2 of [6] by a componentwise application of the standard state augmentation techniques of optimal control theory, and by noting the inclusion

$$\bigcup_{\alpha\geqq 0}\alpha\,\partial d_\Lambda\subset N_\Lambda,$$

which follows from (2.1).

We conclude this section with a result that extends to multiprocesses the fact (familiar in the single process case) that for autonomous problems the Hamiltonian is constant. The problem (P) is autonomous when the functions ϕ_i and L_i have no dependence on t, and when the control set U_t^i is the same set U_0^i for all t.

COROLLARY 3.1. *Under the hypotheses of Theorem* 3.1, *when in addition* (P) *is autonomous, then the conclusions of the theorem can be supplemented by the following. For each* i, *there is a constant* h^i *such that* $h_0^i=h_1^i=h^i$ *and*

$$h^i=\sup_{w\in U_0^i} H_i(x_i(t), w, p_i(t), \lambda) \quad \textit{for all } t \textit{ in } [T_0^i, T_1^i].$$

We shall merely sketch the proof, which parallels closely that of Theorem 3.6.1 of [4]. We introduce new (additional) control variables v_i (joining w_i) and state variables z_i (joining y_i) ($i=1, 2,\cdots, k$), together with the following modified dynamics:

$$\dot{y}_i=(1+v_i)\phi_i(y_i, w_i), \qquad \dot{z}_i=(1+v_i).$$

The control components v_i are constrained to $[-\varepsilon, \varepsilon]$, where ε is a small positive number. With the help of these additional variables we define a new multiprocess problem ($\tilde{\text{P}}$) in terms of states (z_i, y_i) and controls (v_i, w_i) for which each of the k components evolves on the fixed time interval $[T_0^i, T_1^i]$, and in which the quantity

$$f(\{z_i(T_0^i), z_i(T_1^i), y_i(T_0^i), y_i(T_1^i)\})$$

is to be minimized. The corresponding endpoint constraint set $\tilde{\Lambda}$ is given by

$$\tilde{\Lambda}=\{\{\tau_0^i, \tau_1^i, z_0^i, y_0^i, z_1^i, y_1^i\}: \tau_0^i=T_0^i, \tau_1^i=T_1^i, \{z_0^i, z_1^i, y_0^i, y_1^i\}\in\Lambda\}.$$

We note that the state-control components

$$\{\{z_i\equiv t), x_i, (v_i\equiv 0), u_i\}$$

are admissible for the new problem ($\tilde{\text{P}}$). In fact, a standard use of the "Erdmann transformation" (see [4, § 3.6]) shows that (for ε small) this new multiprocess is optimal

for $(\tilde{P})$. On applying Theorem 3.1, we deduce the existence of certain adjoint variables (q_i, p_i), where the q_i are (scalar) constants and where the p_i satisfy the required adjoint equation. The maximization condition is seen to assert that, almost everywhere, the function

$$(v, w) \to (1+v)\{H_i(x_i(t), w, p_i(t), \lambda) + q_i\}$$

is maximized over $[-\varepsilon, \varepsilon] \times U_0^i$ at $(v, w) = (0, u_i(t))$. It follows that for all $t \in [T_0^i, T_1^i]$ we have

$$-q_i = \max_{w \in U_0^i} H_i(x_i(t), w, p_i(t), \lambda),$$

and that the constant h^i in the statement of the corollary may be taken to be $-q_i$. The transversality conditions for $(\tilde{P})$ translate directly to the requisite ones for (P). Finally, we need to confirm that if λ is zero, then $\sum_i \| p_i(T_1^i) \|$ is not. Suppose to the contrary that both vanish. It follows then from the equation above that q_i is zero for each i, contradicting the known fact that (when $\lambda = 0$) $\sum_i \| (p_i(T_1^i), q_i) \|$ is positive.

4. The free time optimal control problem. In the case that the integer $k = 1$, the optimal multiprocess problem can be expressed as follows:

$$\text{minimize} \quad \int_a^b L(t, x(t), u(t))\, dt + f(a, b, x(a), x(b))$$

subject to

(E1)
$$\begin{aligned} &\dot{x}(t) = \phi(t, x(t), u(t)) && \text{a.e. } t \in [a, b], \\ &u(t) \in U_t && \text{a.e. } t \in [a, b], \\ &x(t) \in X_t && \text{for all } t \in [a, b], \\ &(a, b, x(a), x(b)) \in C. \end{aligned}$$

(We have written L in place of L_1, etc.)

Here we recognize a version of the free time optimal control problem. Necessary conditions for solutions to free time optimal control problems date back to the work of Pontryagin and his associates in the 1950s [18], their applicability has been extended by a number of authors [9], [24], [17], [4] and they feature in many expository accounts of optimal control theory. In these optimality conditions, the maximum principle for the fixed time problem is supplemented by a boundary condition on the Hamiltonian function $H := p \cdot \phi - \lambda L$ evaluated along the optimal process $(x^*(\cdot), u^*(\cdot))$ and costate function:

$$t \to \max_{w \in U_t} H(t, x^*(t), w, p(t), \lambda).$$

The hypotheses under which this extra condition has been proved require, at the very least, the Hamiltonian function to be continuous in some sense at the optimal endtimes. (See, e.g., [2] or [7].) Now optimal control problems arise where, possibly as a result of an instantaneous change in an exogeneous input, the data is discontinuous in the time variable. This accounts for developments in fixed time optimal control theory in which the smoothness hypotheses in [18], regarding time dependence, are relaxed to measurability (see, e.g., [2], [9], [4]). The reasons for treating free time problems with data merely measurable in the time variable are as cogent, yet a maximum principle for a general class of problems having this feature has until now been lacking. This omission is remedied by the following theorem. As usual H denotes the Hamiltonian function

$$H(t, x, u, p, \lambda) = p \cdot \phi(t, x, u) - \lambda L(t, x, u).$$

THEOREM 4.1. *Let* $(a^*, b^*, x^*(\cdot), u^*(\cdot))$ *solve problem* (E_1). *Assume the following*:

- *For each* $y \in \mathbb{R}^n$, $\phi(\cdot, y, \cdot)$ *is* $\mathcal{L} \times \mathcal{B}$ *measurable.*
- $U \subset \mathbb{R} \times \mathbb{R}^m$ *is Borel measurable.*

There exists a constant K *such that*

- $|\phi(t, y, w)| \leqq K$ *whenever* $(t, y, w) \in \mathbb{R} \times X_t \times U_t$, *and*
- $|\phi(t, y, w) - \phi(t, y', w)| \leqq K|y - y'|$ *whenever* $(t, y, w), (t, y', w) \in \mathbb{R} \times X_t \times U_t$,

and

$$\text{graph}\,\{x^*(\cdot)\} \subset \text{interior}\,\{X\}.$$

Then there exist real numbers h_0, h_1, *and* λ $(\lambda = 0$ *or* $1)$ *and an absolutely continuous function* $p(\cdot) : [a^*, b^*] \to \mathbb{R}^n$ *such that* $\lambda + |p(b^*)| > 0$,

$$-\dot{p}(t) \in \partial_x H(t, x^*(t), u^*(t), p(t), \lambda) \quad \text{a.e. } t \in [a^*, b^*],$$

$$H(t, x^*(t), u^*(t), p(t), \lambda) = \max_{w \in U_t} H(t, x^*(t), w, p(t), \lambda) \quad \text{a.e. } t \in [a^*, b^*],$$

$$h_0 \in \text{co ess}_{t \to a^*} \left[\sup_{w \in U_t} H(t, x(a^*), w, p(a^*), \lambda) \right],$$

$$h_1 \in \text{co ess}_{t \to b^*} \left[\sup_{w \in U_t} H(t, x(b^*), w, p(b^*), \lambda) \right],$$

$$(-h_0, h_1, p(a^*), -p(b^*)) \in N_C + \lambda\, \partial f,$$

where the normal cone N_C *and the generalized gradient* ∂f *are evaluated at* $(a^*, b^*, x(a^*), x(b^*))$.

Example. A problem in production planning illustrates application of the new maximum principle for free time problems:

$$\text{minimize} \quad -g(T, x(T)) + \int_0^T \alpha(t) u(t)\, dt$$

subject to

$$(E_1)' \qquad \dot{x}(t) = -x(t) + u(t) \quad \text{a.e. } t \in [0, T],$$

$$x(0) = 0,$$

$$u(t) \in [0, 1] \quad \text{a.e. } t \in [0, T].$$

Here $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a given continuously differentiable function and the function α is taken to be

$$\alpha(t) = \begin{cases} 0.5 & \text{if } 0 \leqq t \leqq 2, \\ \bar{\alpha} & \text{if } 2 < t. \end{cases}$$

The termination time T is a choice variable, constrained to satisfy $0 \leqq T \leqq \bar{T}$. ($\bar{T} > 2$ is a given constant.)

An interpretation of this problem is as follows. Money is borrowed and invested in production. The sum borrowed at time t, $u(t)$, is constrained: $0 \leqq u(t) \leqq 1$. The interest rate is $\alpha(t) = 0.5$ up to time $t = 2$, when it jumps to the punitive level $\bar{\alpha}$. $x(t)$ represents the amount of product available for sale at time t. Production is terminated at time $t = T$, and a gross profit $g(T, x(T))$ is realised. We aim to maximize net profit over the class of borrowing policies $\{u(t): 0 \leqq t \leqq T\}$. Net profit is clearly expressible as the negative of the cost function. The purpose of the high interest rate $\bar{\alpha}$ is to discourage borrowing after time $t = 2$. We might ask then how high must $\bar{\alpha}$ be set to make production unprofitable after time $t = 2$? Theorem 4.1 supplies a condition on the value of $\bar{\alpha}$ for $T = 2$ to be an optimal termination time. To state this we define a

process $\{(x(t), u(t)), 0 \leqq t \leqq T\}$ to be an extremal on $[0, T]$ when there exists an absolutely continuous function $p(\cdot)$ on $[0, T]$ such that

$$\begin{aligned}
&\dot{x}(t) = -x(t) + u(t) \quad \text{a.e. } t \in [0, T],\\
&x(0) = 0,\\
&\dot{p}(t) = p(t) \qquad\qquad \text{a.e. } t \in [0, T],\\
&p(T) = g_x(T, x(T)),\\
&u(t) = \begin{cases} 1 & \text{if } p(t) - \alpha(t) > 0 \quad \text{a.e. } t \in [0, T],\\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

In applying Theorem 4.1 we have in the present case

$$f(\tau_0, \tau_1, x_0, x_1) = -g(\tau_1, x_1), \qquad C = \{0\} \times [0, \bar{T}] \times \{0\} \times \mathbb{R}.$$

It is a simple matter to deduce the following proposition.

PROPOSITION 4.2. *Suppose there exists a minimizing process for problem* $(E_1)'$ *with termination time* $T = 2$. *Then there exists an extremal* $\{(x(t), u(t)), 0 \leqq t \leqq 2\}$ *on* $[0, 2]$ *such that*

$$g_x x(2) - g_T \in [0, \infty) \cap [g_x - \bar{\alpha}, g_x - 0.5]. \tag{4.1}$$

In this last inclusion the partial derivatives g_T and g_x are evaluated at $(T, x) = (2, x(2))$. Condition (4.1) is the new information provided by Theorem 4.1 beyond the content of the classical maximum principle; it corresponds to the part of the transversality condition involving the essential value h_1.

It is of interest to examine how closely the inclusion (4.1) captures the range of $\bar{\alpha}$ values for which $T = 2$ is an optimal termination point. To do this we specialize somewhat and consider a particular function $g(T, x)$:

$$g(T, x) = Tx \quad \text{for all } T \geqq 0, \quad x \in \mathbb{R}. \tag{4.2}$$

For each $T \geqq 0$, let us denote by $E_1'(T)$ a variant of $(E_1)'$ in which T is treated as a fixed parameter. For each $T \geqq 0$ we can determine the process that satisfies the necessary conditions for optimality of the fixed time maximum principle (there is just one such process). But $E_1'(T)$ can be shown to have a solution and consequently the process so derived actually solves it. Substituting back into the cost function we obtain a formula for the value function $\eta(T)$, relative to T, namely

$$\eta(T) = \min E_1'(T), \qquad T \geqq 0.$$

We find the following:

(a) If $\bar{\alpha} > 2$ and $|T - 2|$ is small enough

$$\eta(T) = \begin{cases} -Te^{-(T-2)} + 0.5[1 - \ln(0.5/T)] - 0.5(T-2) & \text{for } T > 2,\\ -T + 0.5[1 - \ln(0.5/T)] & \text{for } T \leqq 2. \end{cases}$$

(b) If $\bar{\alpha} \leqq 2$ and $|T - 2|$ is small enough

$$\eta(T) = \begin{cases} -T + 0.5[\ln(0.5/T)] + (\bar{\alpha} - 0.5)(T-2) & \text{for } T > 2,\\ -T + 0.5[1 - \ln(0.5/T)] & \text{for } T \leqq 2. \end{cases}$$

The value function η has left and right derivatives of all orders at $T = 2$. These are easily calculated. Of interest are the following:

$$\begin{aligned}
&D^-\eta(T=2) = -0.75\\
&D^+\eta(T=2) = \begin{cases} -0.75 & \text{if } \bar{\alpha} > 2,\\ -1.25 + \bar{\alpha} & \text{if } \bar{\alpha} \leqq 2, \end{cases}\\
&(D^+)^2\eta(T=2) = -0.25 \quad \text{if } \bar{\alpha} = 1.25.
\end{aligned}$$

It is clear from these values that $T=2$ *is a local minimum of* η *if and only if* $\bar{\alpha}>1.25$.

We return now to the condition on $\bar{\alpha}$ provided by the free time maximum principle. If $\{x(t), u(t); 0\leqq t\leqq 2\}$ is the extremal on $[0, 2]$ of Proposition 4.2, then it is easy to show that $x(2)=0.75$.

The inclusion now becomes $1.5-0.75\in[0,\infty)\cap[2-\bar{\alpha}, 2-0.5]$, i.e., $2-\bar{\alpha}\leqq 0.75\leqq 1.5$ or $\bar{\alpha}\geqq 1.25$.

Comparing this inequality with our findings from calculating the value function, we see that the free time maximum principle provides a rather precise estimate, namely $[1.25,\infty)$, of the set $(1.25,\infty)$ of $\bar{\alpha}$ values for which borrowing a short time after time $t=2$ is unprofitable. It fails to catch the point $\bar{\alpha}=1.25$ since if $\bar{\alpha}=1.25$ profits can be increased by overrunning the time $t=2$ can be ascertained only from second-order properties of the value function, and the maximum principle is a first-order optimality condition.

This example draws attention to the fact that even if data associated with a class of free time problems is discontinuous in the time variable at just one point, instances when this point is the optimal termination time are hardly anomalous. Indeed this is the case for a whole range of values of $\bar{\alpha}$. We see also that the information supplied by the new free time maximum principle can be rather precise in such circumstances.

Suppose there is an optimal process for problem $(E_1)'$, with termination time $T=2$. Limiting attention to the particularly simple terminal cost term (4.2), we find that information present in the free time maximum principle regarding the optimal free time amounts to

$$0\in\partial\eta(T=2), \tag{4.3}$$

i.e., $T=2$ is a stationary point of the value function. We emphasize however that condition (4.1) from the free time maximum principle supplies information about optimal processes even when we cannot derive a convenient formula for the value function and use inclusion (4.3) directly.

5. Impulse control. In this section we derive from Theorem 3.1 a maximum principle for optimal impulse control problems of the following form:

$$\text{minimize}\quad g(a, x(a), \tau, x(\tau^-), x(\tau^+), b, x(b))+\int_a^b L(t, x(t), u(t))\,dt$$

subject to

$$(E_2)\qquad \begin{aligned} &\dot{x}(t)=\phi(t, x(t), u(t)) && \text{a.e. } t\in[a, b],\\ &u(t)\in U_t && \text{a.e. } t\in[a, b],\\ &x(t)\in X_t && \text{for all } t\in[a, b],\\ &(a, x(a), \tau, x(\tau^-), x(\tau^+), b, x(b))\in C. \end{aligned}$$

Here, $L:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$, $\phi:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}^n$ and $g:(\mathbb{R}\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^n)\times(\mathbb{R}\times\mathbb{R}^n)\to\mathbb{R}$ are given functions and $U\subset\mathbb{R}\times\mathbb{R}^m$, $X\subset\mathbb{R}\times\mathbb{R}^m$ and $C\subset\{((t_1, x_1), (t_2, x_2^-, x_2^+), (t_3, x_3))\,|\,t_1\leqq t_2\leqq t_3\}$ are given sets.

Minimization is conducted over elements

$$\{a, \tau, b, x(\cdot):[a, b]\to\mathbb{R}^n, u(\cdot):[a, b]\to\mathbb{R}^m\}$$

in which a, τ and b are numbers that satisfy $a\leqq\tau\leqq b$, $x(\cdot)$ is a function whose restrictions to $[a, \tau)$ and $(\tau, b]$ are absolutely continuous, and $u(\cdot)$ is a measurable

function. The state $x(t)$ is permitted to jump at time $t=\tau$. The departure and arrival states are denoted by $x(\tau^-), x(\tau^+)$:

$$x(\tau^-) := \begin{cases} \lim_{s\uparrow\tau} x(s) & \text{if } \tau > a, \\ x(a) & \text{if } \tau = a, \end{cases}$$

and

$$x(\tau^+) := \begin{cases} \lim_{s\downarrow\tau} x(s) & \text{if } \tau < b, \\ x(b) & \text{if } \tau = b. \end{cases}$$

We write $H = p\cdot\phi - \lambda L$.

THEOREM 5.1. *Let* $\{a^*, \sigma, b^*, x^*(\cdot), u^*(\cdot)\}$ *solve problem* (E_2). *Assume that hypotheses* (H1)-(H4) *of* § 3 *are satisfied when we substitute* ϕ, L, U *and* X *in place of* ϕ^i, L^i U^i *and* X^i. *Suppose also that* g *is locally Lipschitz continuous,* C *is closed, and there exists* $\varepsilon > 0$ *such that*

$$\text{graph}\,\{x^*(\cdot)\} + \varepsilon B \subset X.$$

Then there exist numbers $\alpha, \beta^-, \beta^+, \gamma,$ *and* λ $(\lambda \geqq 0)$ *and a function* $p(\cdot):[a^*, b^*]\to\mathbb{R}^n$ *such that the restrictions of* $p(\cdot)$ *to* $[a^*, \sigma)$ *and* $(\sigma, b^*]$ *are absolutely continuous,* $\lambda + |p(\cdot)|_{L_1} \neq 0$,

$$-\dot{p}(t) \in \partial_x H(t, x^*(t), u^*(t), p(t), \lambda) \quad \text{a.e. } t\in[a^*, b^*],$$

$$H(t, x^*(t), u^*(t), p(t), \lambda) = \max_{w\in U_t} H(t, x^*(t), u^*(t), w, p(t), \lambda) \quad \text{a.e. } t\in[a^*, b^*],$$

$$\alpha \in \text{co}\,\operatorname*{ess}_{t\to a^*} \mathcal{H}(t, x^*(a^*), p(a^*)), \qquad \beta^- \in \text{co}\,\operatorname*{ess}_{t\to\sigma} \mathcal{H}(t, x^*(\sigma^-), p(\sigma^-)),$$

$$\beta^+ \in \text{co}\,\operatorname*{ess}_{t\to\sigma} \mathcal{H}(t, x^*(\sigma^+), p(\sigma^+)), \qquad \gamma \in \text{co}\,\operatorname*{ess}_{t\to b^*} \mathcal{H}(t, x^*(b^*), p(b^*)),$$

$$(-\alpha, p(a^*), -(\beta^+ - \beta^-), -p(\sigma^-), p(\sigma^+), \gamma, -p(b^*)) \in N_C + \lambda\,\partial g.$$

Here $\mathcal{H}(t, x, p) := \sup_{w\in U_t} H(t, x, w, p, \lambda)$. *The normal cone* N_C *and generalized gradient* ∂g *are evaluated at*

$$(a^*, x^*(a^*), \sigma, x^*(\sigma^-), x^*(\sigma^+), b^*, x^*(b^*)).$$

Observe that problem (E_2) can be reformulated as an optimal multiprocess problem as in § 3. Two component processes are involved, associated with the restrictions of the control function and state trajectory to either side of the jump time τ. The dynamics and state and control constraints are the same as those in problem (E_2) for both component processes. The fact that the component processes are concatenated is taken account of by choosing the set Λ in problem (P) of § 3 to be

$$\Lambda = \{\{\tau_0^i, \tau_1^i, x_0^i, x_1^i\} \mid \tau_1^1 = \tau_0^2 \text{ and } (\tau_0^1, x_0^1, \tau_1^1, x_1^1, x_0^2, \tau_1^2, x_1^2) \in C\}. \tag{5.1}$$

The assertions of Theorem 5.1 follow now from application of Theorem 3.1, in which the optimal multiprocess considered is that associated with $\{a^*, \sigma, b^*, x^*(\cdot), u^*(\cdot)\}$. Bearing in mind the special structure of Λ (see (5.1)), we obtain the transversality condition by an application of Proposition 2.1(i), and by observing that the function g in the cost function can be assumed to have values that do not depend on the τ_1^1 variable.

Let us consider a special case, where the conditions in Theorem 5.1 simplify considerably. The problem in question is one where the underlying time interval $[a, b]$, is fixed and where it is the jump vector, $x(\tau^+) - x(\tau^-)$, rather than the departure and

arrival states of the jump, $x(\tau^+)$ and $x(\tau^-)$, which feature explicitly in the cost and constraints:

$$\text{minimize} \qquad g_1(x(a), x(b)) + \mathscr{K}(\tau, x(\tau^+) - x(\tau^-)) + \int_a^b L(t, x(t), u(t))\, dt$$

$$(E_3) \quad \text{subject to } \dot{x} = \phi, \quad u(t) \in U_t, x(t) \in X_t,$$

$$(x(a), x(b)) \in C_1, (\tau, x(\tau^+) - x(\tau^-)) \in C_2, \qquad a \leqq \tau \leqq b.$$

Let $\{\tau = \sigma, x^*(\cdot), u^*(\cdot)\}$ solve problem (E_3). It is assumed that the data of problem (E_3) satisfies the hypotheses of Theorem 5.1, when viewed as a special case of problem (E_2).

Applying Theorem 5.1, and taking note of Proposition 2.1(ii), we can show the following corollary.

COROLLARY 5.2. *Take* $\{\tau = \sigma, x^*(\cdot), u^*(\cdot)\}$ *as above. Then, under the stated hypotheses, there exists an absolutely continuous function* $p(\cdot): [a, b] \to \mathbb{R}^n$ *and* $\lambda \geqq 0$ $(\lambda + \| p(1) \| \neq 0)$ *such that*

$$(5.2) \qquad -\dot{p}(t) \in \partial_x H(t, x^*(t), u^*(t), p(t), \lambda) \quad \text{a.e.},$$

$$(5.3) \qquad H(t, x^*(t), u^*(t), p(t), \lambda) = \max_{v \in U_t} H(t, x^*(t), v, p(t), \lambda) \quad \text{a.e.},$$

$$(5.4) \qquad (p(a), -p(b)) \in N_{C_1}(x^*(a), x^*(b)) + \lambda\, \partial_{g_1}(x^*(a), x^*(b)),$$

$$(5.5) \qquad (\beta^+ - \beta^-, p(\sigma)) \in N_{C_2}(\sigma, x^*(\sigma^+) - x^*(\sigma^-)) + \lambda\, \partial \mathscr{K}(\sigma, x^*(\sigma^+) - x^*(\sigma^-))$$

where

$$\beta^- \in \operatorname{co} \operatorname*{ess}_{t \to \sigma} \mathscr{h}(t, x^*(\sigma^-), p(\sigma)) \quad \text{and} \quad \beta^+ \in \operatorname{co} \operatorname*{ess}_{t \to \sigma} \mathscr{h}(t, x^*(\sigma^+), p(\sigma)).$$

(H and $\mathscr{h}$ were defined in, and preceding, the statement of Theorem 5.1).

We recognize in these optimality conditions the assertions of the standard maximum principle for fixed time problems involving an absolutely continuous costate function $p(\cdot): [a, b] \to \mathbb{R}^n$, namely (5.2)-(5.4), supplemented by certain information concerning the jump, namely (5.5). In the case that the function $\mathscr{h}$ is continuous, $C_2 = \mathbb{R} \times \mathbb{R}^n$ and $\mathscr{K} \equiv 0$, for instance, this takes the form

$$p(\sigma) = 0$$

and

$$\sup_w H(\sigma, p(\sigma), x^*(\sigma^+), w, \lambda) = \sup_w H(\sigma, p(\sigma), x^*(\sigma^-), w, \lambda).$$

For simplicity of exposition so far we have considered impulse control problems involving at most one jump. As a simple corollary of Theorem 5.1 we also obtain optimality conditions for a problem similar to (E_2), but where we permit jumps at (at most) k times τ_i, $i = 1, \cdots, k$. These jump times are choice variables that, along with the endpoints of the state trajectories and their values near the jump times, are constrained according to

$$(a, x(a), \{\tau_i, x(\tau_i^-), x(\tau_i^+)\}, b, x(b)) \in \tilde{C}$$

and enter into the cost through a term:

$$\tilde{g}(a, x(a), \{\tau_i, x(\tau_i^-), x(\tau_i^+)\}, b, x(b)).$$

(Note that cases where there are less than k jumps are accommodated since we permit $\tau_i^- = \tau_i^+$ for some i.) We will call such problems general impulse control problems. Let $(a = a^*, b = b^*, \{\tau_i = \sigma_i\}, x(\cdot), u(\cdot))$ solve the general impulse control problem. Then conditions are satisfied similar to those in Theorem 5.1, except that now the costate function $p(\cdot)$ may jump at times τ_i, $i = 1, \cdots, k$ and the transversality condition takes the form

$$(-\alpha, p(a^*), \{-(\beta_i^+ - \beta_i^-), -p(\sigma_i^-), p(\sigma_i^+)\}, \gamma, -p(b^*)) \in N_{\tilde{C}} + \lambda\, \partial \tilde{g}$$

in which

$$\beta_i^- \in \operatorname{co} \operatorname*{ess}_{t \to \sigma_i} \mathcal{H}(t, x^*(\sigma_i^-), p(\sigma_i^-)) \quad \text{and} \quad \beta_i^+ \in \operatorname{co} \operatorname*{ess}_{t \to \sigma_i} \mathcal{H}(t, x^*(\sigma_i^+), p(\sigma_i^+)), \qquad i = 1, \cdots, k.$$

Impulse control problems of the type studied here arise in, for example, operations research and production planning (see [10]) and have been the subject of earlier research. Previous work has been directed for the most part at characterization of optimal processes in terms of solutions to a Hamilton-Jacobi equation [1] and [10]. By contrast, we provide *necessary conditions* of optimality in the form of a maximum principle.

It is true that there is an existing necessary conditions literature for optimal control problems where the state trajectories are permitted to be functions of bounded variation. This is an outgrowth of certain problems in flight mechanics and investment planning. The state trajectories in question are associated with generalized control functions that are measures (see, e.g., [19], [20], [23] and [24]). The formulation employed in such work is suited to problems where the number of jumps is unbounded and the cost and constraints are expressed in terms of such quantities as the total variation of the jumps. If specialized to apply to the kinds of problems addressed by Theorem 5.1 (and its multiple jump analogues), stringent additional hypotheses need to be made. We require the a priori information that the constraint on the number of jumps is nonbinding and the measures involved has no continuous singular component, and we also require that the cost on the jumps is expressible in terms of a positively homogeneous function of the jump vector.

Derivations of certain special cases of Theorem 5.1 and of its multiple jump counterpart, together with some heuristic calculations, are to be found in [21] and [22] and the references therein. Recently Frankowska [8] has given necessary conditions for a class of multiple jump impulse control problems, in which the data is assumed continuous in the time variable.

6. Laws of reflection and refraction. Fermat's principle in optics identifies the path of a ray of light with an arc, joining the endpoints in question, along which light travels in minimum time. In the case of reflection, the minimum is with respect to arcs which visit the reflecting surface.

It is well known (see, e.g., [12]) that Fermat's principle predicts Snell's laws of reflection and refraction; namely, for reflection,

$$\theta_1 = \theta_2,$$

where θ_1 is the angle of incidence and θ_2 is the angle of reflection, and for refraction,

$$g_1 \sin \theta_1 = g_2 \sin \theta_2$$

where g_1 is the refractive index of the medium of the incident ray, θ_1 is the angle of incidence, g_2 is the refractive index of the medium of the refracted ray and θ_2 is the angle of refraction.

The traditional approach is to obtain a formula for the path that satisfies Fermat's principle and then to check Snell's laws directly. The minimization is carried out in two phases. Initially the point of incidence is fixed at an arbitrary value. The transit time is subsequently minimized over points of incidence. This approach, dependent as it is on obtaining a formula for the minimizing path, is very restrictive as far as permitted geometry of the boundary and variation of refractive index in the media is concerned.

The problem of predicting Snell's laws fits neatly into optimal multiprocess theory. We have here a dynamic optimization problem that breaks into two regimes connected at the point of incidence.

We shall find that optimal multiprocess theory predicts a version of Snell's law for a very large class of inhomogeneous media, and for interfaces Σ which are required to be merely closed sets.

The following data is given:

$$\text{functions } g_1 : \mathbb{R}^n \to \mathbb{R}, \qquad g_2 : \mathbb{R}^n \to \mathbb{R},$$

a closed set Σ and points x_0, x_1 in $\mathbb{R}^n$. It is assumed that g_1 and g_2 are locally Lipschitz continuous, and have values everywhere greater than zero.

Consider the following optimization problem:

$$\text{minimize } \int_0^\tau g_1(y(\sigma))\, d\sigma + \int_\tau^T g_2(y(\sigma))\, d\sigma$$

$$\text{subject to } 0 \leqq \tau \leqq T, \tag{E_4}$$

$$\dot{y}(\sigma) \in S \quad \text{a.e. } \sigma \in [0, T],$$

$$y(0) = x_0, \quad y(T) = x_1, \quad y(\tau) \in \Sigma.$$

Here the numbers τ and T, and the Lipschitz continuous function $y : [0, T] \to \mathbb{R}^n$ are the choice variables. S is the surface of the unit ball in $\mathbb{R}^n$.

In this problem we interpret the independent variable σ as arclength. Keeping in mind that the refractive index is the reciprocal of the speed of light, we recognize the cost function as the transit time through two adjoining media. The constraint "$\dot{y}(\sigma) \in S$" simply corresponds to the well-known relationship between increments of the coordinate values and arclength along a curve, namely

$$\sum_i \left(\frac{dy_i}{d\sigma} \right)^2 = 1.$$

A version of Fermat's principle then is that the path of a ray of light is a minimizer for this problem. There is the implicit assumption in our formulation that if we fix the point of incidence at a minimizing value, then the segments of the path on the first and second sides are governed entirely by g_1 and g_2, respectively.

THEOREM 6.1. *Let* $((y : [0, T] \to \mathbb{R}^n), \tau, T)$ *be a solution to* (E_4) *and suppose that* $0 \leqq \tau < T$. *Then* $y(\cdot)$ *is continuously differentiable on* $[0, \tau)$ *and* (τ, T), *and* $\dot{y}(\cdot)$ *has left and right limits* $\dot{y}(\tau^-)$ *and* $\dot{y}(\tau^+)$ *at* τ. *We have, either*

$$g_1(y(\tau)) = g_2(y(\tau)) \quad \textit{and} \quad \dot{y}(\tau^-) = \dot{y}(\tau^+),$$

or there exists some nonzero vector $d \in N_\Sigma(y(\tau))$ *such that both conditions* (i) *and* (ii) *below are satisfied*:

(i) $$\dot{y}(\tau^+) \in \text{span}\,\{\dot{y}(\tau^-), d\},$$

(ii) $$g_1(y(\tau)) \sin \theta_1 = g_2(y(\tau)) \sin \theta_2,$$

where

$$\theta_1 = \cos^{-1}\left(\frac{d \cdot \dot{y}(\tau^-)}{\|d\|\,\|\dot{y}(\tau^-)\|}\right) \quad \textit{and} \quad \theta_2 = \cos^{-1}\left(\frac{d \cdot \dot{y}(\tau^+)}{\|d\|\,\|\dot{y}(\tau^+)\|}\right).$$

In this theorem, θ_1 (respectively, θ_2) will be recognized as the angle between the tangent vector to the incoming (respectively, outgoing) ray at the point of incidence and the normal vector d.

When we treat refraction, we view $g_1(g_2)$ as the refractive index in the interior of the medium through which the incident (refracted) ray travels. The theorem also covers the phenomenon of reflection. Here we set $g_1 = g_2 (=: g)$. Since g is positive-valued, condition (ii) simply becomes

$$\theta_1 = \theta_2.$$

Proof. We may identify (E_4) with a multiprocess problem in which the two components are the state arc $y(\cdot)$ of (E_4) restricted to $[0, \tau]$ and to $[\tau, T]$, and in which Λ is the set

$$\{[0, \tau, x_0, y, \tau, T, y, x_1]: 0 \leqq \tau \leqq T, y \in \Sigma\}.$$

Now apply the corollary to Theorem 3.1. Since the component processes are concatenated, the necessary conditions can be considered as expressed in terms of a cost multiplier $\lambda \geqq 0$ and a *single* function $p(\cdot):[0, T] \to \mathbb{R}^n$, both of which do not vanish. We see that the following conditions are satisfied:

$$\begin{aligned} &\|p(t)\| - \lambda g_1(y(t)) = 0, \qquad t \leqq t < \tau, \\ &\|p(t)\| - \lambda g_2(y(t)) = 0, \qquad \tau < t \leqq T, \end{aligned} \tag{6.1}$$

$$p(t) \cdot u(t) = \max\{p(t) \cdot v \mid \|v\| = 1\} \quad \text{a.e. } t \in [0, T] \tag{6.2}$$

(both (6.1) and (6.2) follow from the "maximization of the Hamiltonian" condition and the fact that h_1^2 is zero)

$$((-p(\tau^-), p(\tau^+)) \in N_{\tilde{\Sigma}}(y(\tau), y(\tau)) \tag{6.3}$$

(the transversality condition). Here $p(\tau^-)$ and $p(\tau^+)$ are the limits of the function $p(\cdot)$ from the left and the right, respectively. The set $\tilde{\Sigma}$ is

$$\tilde{\Sigma} := \{(a, a): a \in \Sigma\}.$$

We must have $\lambda \neq 0$ since otherwise $p(\cdot) \equiv 0$ by (6.1) in violation of the nontriviality of the multipliers. Then we are permitted to set $\lambda = 1$. By hypothesis, $g_1, g_2 \geqq \varepsilon$ for some $\varepsilon > 0$. In view of (6.1), $\|p(t)\| \geqq \varepsilon$ a.e. By continuity then, $p(\tau^-), p(\tau^+)$ and $p(t), t \in [0, T] \backslash \{\tau\}$ are all nonzero.

From (6.2)

$$u(t) = p(t)/\|p(t)\| \quad \text{a.e. } t \in [0, T].$$

Since the function on the right-hand side is continuous at all points in $[0, T] \backslash \tau$ and has left and right limits at τ, and since $y(\cdot)$ is Lipschitz continuous and satisfies $\dot{y}(t) = u(t)$ almost everywhere, it follows that $y(\cdot)$ is continuously differentiable on $[0, \tau)$ and $(\tau, T]$, and that $\dot{y}(\cdot)$ has left and right limits, $\dot{y}(\tau^-)$ and $\dot{y}(\tau^+)$, at τ given by

$$\dot{y}(\tau^-) = \frac{p(\tau^-)}{\|p(\tau^-)\|} \quad \text{and} \quad \dot{y}(\tau^+) = \frac{p(\tau^+)}{\|p(\tau^+)\|}. \tag{6.4}$$

By (6.3) and Proposition 2.1(i) however

$$-p(\tau^-)+p(\tau^+)=d \tag{6.5}$$

for some $d \in N_\Sigma(y(\tau))$.

We deduce from (6.1) and the continuity properties of y, p, g_1 and g_2 that

$$\|p(\tau^-)\| = g_1(y(\tau)) \quad \text{and} \quad \|p(\tau^+)\| = g_2(y(\tau)).$$

It now follows from (6.4) and (6.5) that

$$g_2(y(\tau))\dot{y}(\tau^+) - g_1(y(\tau))\dot{y}(\tau^-) = d. \tag{6.6}$$

Since $g_2 > 0$ by hypothesis, we deduce that

$$\dot{y}(\tau^+) \in \text{span}\,\{\dot{y}(\tau^-), d\}.$$

It is now convenient to consider two cases.

First suppose $d = 0$. Taking norms across (6.6) we deduce $g_2(y(\tau)) = g_1(y(\tau))$. A further appeal to (6.6) yields the information $\dot{y}(\tau^+) = \dot{y}(\tau^-)$, in accordance with the theorem.

The remaining case is $d \neq 0$. We deduce from (6.6) that

$$g_1(y(\tau))\left[\dot{y}(\tau^-) - \frac{(\dot{y}(\tau)\cdot d)}{\|d\|^2}d\right] = g_2(y(\tau))\left[\dot{y}(\tau^+) - \frac{(\dot{y}(\tau^+)\cdot d)}{\|d\|^2}d\right],$$

i.e.,

$$g_1(y(\tau))\pi^- = g_2(y(\tau))\pi^+ \tag{6.7}$$

where π^- and π^+ are the projections of the vectors $\dot{y}(\tau^-)$ and $\dot{y}(\tau^+)$, respectively, onto the subspace in $\mathbb{R}^n$ orthogonal to the vector d. Since $\dot{y}(\tau^-)$ and $\dot{y}(\tau^+)$ have unit length

$$\|\pi^-\| = \sin\theta_1 \quad \text{and} \quad \|\pi^+\| = \sin\theta_2$$

where

$$\theta_1 = \cos^{-1}\frac{\dot{y}(\tau^-)\cdot d}{\|d\|\,\|\dot{y}(\tau^-)\|} \quad \text{and} \quad \theta_2 = \cos^{-1}\frac{\dot{y}(\tau^+)\cdot d}{\|d\|\,\|\dot{y}(\tau^+)\|}.$$

These identities and (6.7) combine to complete the proof of the theorem.

7. Minimum time control for a robot arm. We pursue our investigations with a detailed analysis of a dynamic optimization problem in which the dynamical equations change at some time τ, where τ is a choice variable. The example is representative of a number of optimization problems having this feature that arise, for example, in the area of multistage rocket control [11], optimal investment [21], and plasma control [22]. We have chosen a relatively simple problem, and one for which an explicit solution may be found by other means, simply to illustrate the nature of the information that the theory of optimal multiprocesses supplies. The example of the next section (taken from renewable resource theory) is one in which explicit two-stage optimization is not available.

The problem is to control a robot arm so that it transfers an object from one location to another as quickly as possible. The mass of the object is comparable with that of the arm, with the result that different dynamical equations apply depending on whether or not the object is being carried.

We adopt a simple mathematical model, involving one spatial variable. The unit (the arm or the arm carrying the load) is initially at rest at the origin. It must be guided to location $x = L$, where it changes mass, and then back to rest at the origin again. A precise formulation is as follows (as usual τ^- and τ^+ will denote limits from the left and right at the point τ):

$$\text{minimize} \quad T$$

$$\text{subject to}$$

$$\ddot{x}(t) = \begin{cases} m_1^{-1}u(t), & 0 \leqq t < \tau \\ m_2^{-1}u(t), & \tau < t \leqq T \end{cases} \quad \text{a.e. } t \in [0, T],$$

$$u(t) \in [-1; +1] \quad \text{a.e. } t \in [0, T], \tag{E$_5$}$$

$$0 \leqq \tau \leqq T,$$

$$\dot{x}(\tau^+) = K\dot{x}(\tau^-),$$

$$x(\tau) = L,$$

$$x(0) = \dot{x}(0) = x(T) = \dot{x}(T) = 0.$$

Minimization is conducted over control policies that, in this instance, are taken to mean triples $(u(\cdot), \tau, T)$ in which τ, T are nonnegative numbers such that $0 \leqq \tau \leqq T$ and $u(\cdot)$ is a measurable function on $[0, T]$ taking values almost everywhere in $[-1, +1]$. The solution of the differential equation that satisfies the left-hand boundary condition is called the corresponding state trajectory. In this program, K, m_1, m_2 and L are positive constants. We shall treat the general problem but special cases of interest are the following:

Dropoff. $m_1 > m_2$, $K = 1$. A load is carried to location $x = L$ and dropped.

Hard pickup. $m_1 < m_2$, $K = m_1/m_2$. A load is picked up at location $x = L$. The load is at rest prior to pick up and the instantaneous change in velocity v of the unit is governed by the principle of conservation of momentum,

$$m_1 v(\tau^-) = m_2 v(\tau^+).$$

Soft pickup. $m_1 < m_2$, $K = 1$. This is the same as the hard pickup case, except that the load may be moving prior to pickup and we are free to choose its velocity at pickup, provided it is the same as that of the robot arm at that instant.

Concerning the solution of this problem, we have Proposition 7.1.

PROPOSITION 7.1. *There is a unique minimizing control policy* $(u^*(\cdot), \tau, T)$ *for problem* (E$_5$). *Let* $v = \dot{x}^*(\tau^-)$, *where* $x^*(\cdot)$ *is the state trajectory corresponding to this control policy. Then* v *is the unique solution of the equation*

$$\frac{m_1^2 v}{(m_1 L + \frac{1}{2}m_1^2 v^2)^{1/2}} + \frac{(Km_2)^2 v}{(m_1 L + \frac{1}{2}m_2^2 K^2 v^2)^{1/2}} = m_1 - Km_2, \tag{7.1}$$

and $(u^*(\cdot), \tau, T)$ *is expressed in terms of* v *according to*

$$T = (Km_2 - m_1)v + 2(m_1 L + \tfrac{1}{2}m_1^2 v^2)^{1/2} + 2(m_2 L + \tfrac{1}{2}m_2^2 K^2 v^2)^{1/2}, \tag{7.2}$$

$$\tau = -m_1 v + 2(m_1 L + \tfrac{1}{2}m_1^2 v^2)^{1/2}, \tag{7.3}$$

$$u^*(t) = \begin{cases} +1, & 0 \leqq t \leqq t_1. \\ -1, & t_1 < t < t_2, \\ +1, & t_2 \leqq t \leqq T. \end{cases} \tag{7.4}$$

In the formula for $u^*(\cdot)$, the switching times, t_1 and t_2, are

$$t_1 = (m_1 L + \tfrac{1}{2} m_1^2 v^2)^{1/2}, \tag{7.5}$$

$$t_2 = (Km_2 - m_1)v + 2(m_1 L + \tfrac{1}{2} m_1^2 v^2)^{1/2} + (m_2 L + \tfrac{1}{2} m_2^2 K^2 v^2)^{1/2}. \tag{7.6}$$

We note that, in the dropoff case, the object is dropped on the outward journey, when the velocity of the arm is positive. In the hard pickup case, the arm has zero velocity at pick up (thus, while impact with the object is permitted, it does not occur). Finally, for soft pickup, the arm picks up the object on the return journey to its starting position, when the velocity is negative.

Proof of Proposition 7.1. Following introduction of the state vector y with components (x, v), where x and v are the position and velocity of the arm, the problem takes the following form:

$$\text{minimize} \quad T$$

$$\text{subject to}$$

$$\dot{y} = \begin{cases} Ay(t) + b_1 u(t), & \text{a.e. } t \in [0, \tau), \\ Ay(t) + b_2 u(t), & \text{a.e. } t \in [\tau, T], \end{cases}$$

$$u(t) \in [-1, +1] \quad \text{a.e. } t \in [0, T], \tag{E_5)$'$}$$

$$0 \leqq \tau \leqq T,$$

$$x(0) = v(0) = x(T) = v(T) = 0,$$

$$x(\tau^+) = x(\tau^-),$$

$$v(\tau^+) = Kv(\tau^-).$$

Here,

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 0 \\ m_1^{-1} \end{bmatrix}, \quad b_2 = \begin{bmatrix} 0 \\ m_2^{-1} \end{bmatrix}.$$

The problem is readily expressible as an optimal multiprocess problem, in which the component control functions are obtained by restricting the function $u(\cdot)$ in problem $(E_5)'$ to the intervals $[0, \tau]$ and $[\tau, T]$, and so on. Let $\{u^*(\cdot), \tau, T\}$ be an optimal control policy, and let $y^*(\cdot) = (x^*(\cdot), v^*(\cdot))$ be the corresponding trajectory. Enlisting the help of Proposition 2.1(i), in computing the normal cone, we make the following deductions from Theorem 3.1. There exist scalar valued functions $p(\cdot)$ and $q(\cdot)$ and a number $\lambda \geqq 0$, not all zero, such that the restrictions of $p(\cdot)$ and $q(\cdot)$ to the intervals $[0, \tau)$ and $(\tau, T]$ are Lipschitz continuous,

$$q(t)u^*(t) = \max_{u \in [-1, +1]} q(t)u, \quad \text{a.e. } t \in [0, T], \tag{7.7}$$

$$\dot{p}(t) = 0, \ \dot{q}(t) = -p(t) \qquad \text{a.e. } t \in [0, T],$$

$$p(\tau^-)v(\tau^-) + m_1^{-1}|q(\tau^-)| = p(\tau^+)v(\tau^+) + m_2^{-1}|q(\tau^+)|,$$

$$q(\tau^+) = K^{-1} q(\tau^-), \tag{7.8}$$

$$m_1^{-1}|q(T)| = \lambda.$$

Before we move on to examine the implications of these conditions, it is convenient to draw attention to certain features of feasible controls, i.e., ones which together with their corresponding trajectories satisfy the constraints of the problem.

We say a control policy $(u(\cdot), \tau, T)$ is a bang-bang policy (with k switches) if there is a finite partition $0\ (=\sigma_0) < \sigma_1 < \cdots < \sigma_k < T(=\sigma_{k+1})$ of the interval $[0, T]$ such that $u(\cdot)$ restricted to $[\sigma_i, \sigma_{i+1}]$ either takes value $+1$ almost everywhere, or takes value -1 almost everywhere, for $i = 0, \cdots, k$. The σ_i's are the switching times. A little study of the phase plane portraits for $(x, \dot{x})$ corresponding to $u(\cdot) \equiv 1$ and to $u(\cdot) \equiv -1$ reveals the following lemma.

LEMMA. (i) *For any feasible bang-bang control policy there must be at least two switching times.*

(ii) *For any feasible bang-bang control policy with two switching times, the control function must assume values in the sequence* $+1, -1, +1$.

Returning now to the necessary conditions, we see that $p(\cdot)$ must be a piecewise constant function and $q(\cdot)$ a piecewise affine function. For both functions the discontinuity (in value or derivative) occurs at time τ.

CLAIM. $q(\tau^-) \neq 0$. Suppose to the contrary that $q(\tau^-) = 0$. We can exclude the possibility that $q(\cdot) \equiv 0$, for then $p(\cdot) \equiv 0$ and $\lambda = 0$, in contradiction to the nontriviality of the multipliers. It follows either $q(\cdot)$ is nonzero on $[0, \tau)$ or on $(\tau, T]$. Suppose first that $q(\cdot)$ is nonzero on $(\tau, T]$. Then $p(\tau^+) \neq 0$. Also, by (7.7), $u^*(\cdot)$ is constant on $(\tau, T]$. It follows that $v(\tau^+) \neq 0$, for clearly the point $(x, v) = (L, 0)$ in phase space cannot be driven to $(x, v) = (0, 0)$ by a constant control. But by (7.8), and since $q(\tau^-) = q(\tau^+) = 0$, we have $p(\tau^-) \neq 0$. This means that $q(\cdot)$ is also nonzero on $[0, \tau)$. But then by (7.7) $u^*(\cdot)$ is constant on $[0, \tau)$. We conclude that $(u^*(\cdot), \tau, T)$ is a feasible bang-bang control policy with at most one switch. This is impossible by the lemma. Similar reasoning excludes the possibility that $q(\cdot)$ is nonzero on $[0, \tau)$. The claim is substantiated.

In view of (7.7) and part (i) of the lemma, $q(\cdot)$ must take value zero at two interior points. There are then two possible configurations for $q(\cdot)$. One corresponds to a bang-bang control sequence $-1, +1, -1$, and is ruled out by the second part of the lemma. The remaining configuration is illustrated in Fig. 1.

We conclude from Fig. 1 that $u^*(\cdot)$ takes the form

$$u^*(t) = \begin{cases} +1 & \text{a.e. } t \in [0, t_1], \\ -1 & \text{a.e. } t \in (t_1, t_2), \\ +1 & \text{a.e. } t \in [t_2, T], \end{cases} \tag{7.9}$$

for some numbers t_1, t_2 satisfying

$$0 < t_1 < \tau < t_2 < T.$$

These are the zero crossing points in Fig. 1.

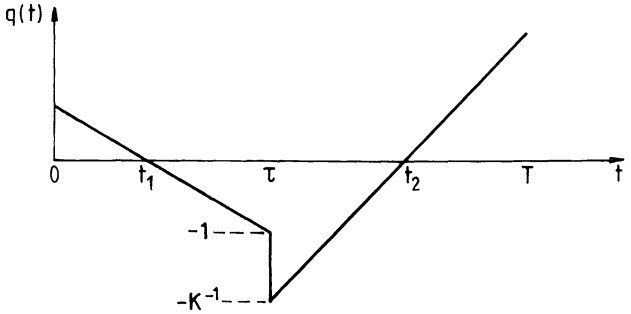

FIG. 1. *Robot arm problem: graph of the costate function* $q(\cdot)$.

Now write v for $v(\tau^-)$. On the interval $[0, \tau]$ the control function $u^*(\cdot)$ drives the unit, which has mass m_1, from the origin to the point $(x = L, \dot{x} = v)$ in the phase plane. Then on the interval $(\tau, T]$ it drives the unit, now with mass m_2, from $(x = L, \dot{x} = Kv)$ back to the origin. A straightforward analysis of phase plane portraits for responses to $m_1\ddot{x} = \pm 1$ and $m_2\ddot{x} = \pm 1$ reveals that

$$-(2L/m_2^2)^{1/2} < v < +(2L/m_1^2)^{1/2}$$

and t_1, t_2, τ and T are determined by the formulae (7.2)-(7.6), once v is found.

What do the optimality conditions say about v? We deduce from the graph of $q(\cdot)$ that

$$q(\tau^-) = -1,\ q(\tau^+) = -K^{-1},$$

$$p(\tau^-)(= -\dot{q}(\tau^-)) = (\tau - t_1)^{-1} = [(m_1 L + \tfrac{1}{2}m_1^2 v^2)^{1/2} - m_1 v]^{-1},$$

$$p(\tau^+)(= -\dot{q}(\tau^+)) = -K^{-1}(t_2 - \tau)^{-1} = K^{-1}[(m_2 L + \tfrac{1}{2}m_2^2 K^2 v^2)^{1/2} + m_2 K v]^{-1}.$$

It follows now from (7.8) that

$$\frac{v}{(m_1 L + \frac{1}{2}m_1^2 v^2)^{1/2} - m_1 v} + \frac{v}{(m_2 L + \frac{1}{2}m_2^2 K^2 v^2)^{1/2} + m_2 K v} = (Km_2)^{-1} - m_1^{-1}.$$

Multiplying across the equation by $Km_1 m_2(\mathscr{L}_1 - m_1 v)(\mathscr{L}_2 + m_2 K v)$ where

$$\mathscr{L}_1 = (m_1 L + \tfrac{1}{2}m_1^2 v^2)^{1/2} \quad \text{and} \quad \mathscr{L}_2 = (m_2 L + \tfrac{1}{2}m_2^2 K^2 v^2)^{1/2}$$

and cancelling terms, we obtain the equation

$$((Km_2)^2 \mathscr{L}_1 + m_1^2 \mathscr{L}_2)v = (m_1 - Km_2)\mathscr{L}_1 \mathscr{L}_2,$$

which implies

$$\frac{m_1^2 v}{(m_1 L + \frac{1}{2}m_1^2 v^2)^{1/2}} + \frac{(Km_2)^2 v}{(m_2 L + \frac{1}{2}m_2^2 K^2 v^2)^{1/2}} = m_1 - m_2 K.$$

This is (7.1).

But the left-hand side of (7.1) defines a function of v which is strictly monotone increasing and assumes all values in the interval

$$(-\sqrt{2}(m_1 + Km_2), \sqrt{2}(m_1 + Km_2)).$$

There is therefore a unique number v satisfying the equation.

We have shown that there is at most one control policy satisfying the necessary conditions, and such a control policy is determined by the formulae in the proposition. We deduce from the existence of feasible control policies (as defined by (7.2)-(7.6) when $v = 0$, for instance) and standard compactness arguments that an optimal control policy exists. There must then be a unique optimal control policy, and it is the one satisfying the necessary conditions. The proposition is proved. For purposes of comparison, we sketch a solution to problem (E_5) by traditional techniques. With every $w \in \mathbb{R}$ we associate an optimization problem $E_5(w)$, a modification of problem (E_5) in which we add a constraint

$$\text{“}x(\tau^-) = w.\text{”}$$

Let $J(\cdot)$ denote corresponding value function,

$$J(w) := \inf E_5(w) \quad \text{for all } w \in \mathbb{R}.$$

For each $w \in \mathbb{R}$, problem $E_5(w)$ decouples into two standard optimal control problems ("time optimal transfer of a double integrator plant with amplitude bounded controls between two fixed points in the phase plane"). A traditional approach is to find v (the minimizing velocity) such that

$$J(v) \leqq J(w) \quad \text{for all } w \in \mathbb{R}$$

and then to determine the optimal control policy as a solution to $E_5(v)$.

Define $\tilde{J}: \mathbb{R} \to \mathbb{R}$ to be

$$\tilde{J}(w) := (m_2K - m_1)w + 2(m_1L + \tfrac{1}{2}m_1^2w^2)^{1/2} + 2(m_2L + \tfrac{1}{2}m_2^2K^2w^2)^{1/2}.$$

Some lengthy but routine calculations along the lines of those in [14] yield the following explicit formula for $J(\cdot)$:

$$J(w) = \begin{cases} (m_2K + m_1)w + 2(-m_1L + \tfrac{1}{2}m_1^2w^2)^{1/2} + 2(m_2L + \tfrac{1}{2}m_2^2K^2w^2)^{1/2} & \text{if } w \geqq \bar{V}, \\ \tilde{J}(w) \quad \text{if } \underline{V} < w < +\bar{V}, \\ -(m_2K - m_1)w + 2(m_1L + \tfrac{1}{2}m_1^2w^2)^{1/2} + 2(-m_2L + \tfrac{1}{2}m_2^2K^2w^2)^{1/2} & \text{if } w \leqq \underline{V}. \end{cases}$$

Here,

$$\bar{V} := (2L/m_1)^{1/2} \quad \text{and} \quad \underline{V} := -(2L/K^2m_2)^{1/2}.$$

We mention that for each $w \in \mathbb{R}$, $E_5(w)$ has a unique optimal control policy, which is a bang-bang policy. For $w \in (\underline{V}, \bar{V})$ the values of the control functions involved switch twice, and for $w \notin [\underline{V}, \bar{V}]$ they switch three times. Relevant phase trajectories are shown in Fig. 2. Note that in solving the problem through multiprocess theory, it was unnecessary to consider controls involving more than two switches.

It can be shown that

$$J(w) \geqq \tilde{J}(w) \quad \text{for all } w \in \mathbb{R}. \tag{7.10}$$

Of course we have

$$J(w) = \tilde{J}(w) \quad \text{for } w \in (\underline{V}, \bar{V}). \tag{7.11}$$

Now it is easy to show that $\tilde{J}(\cdot)$ is a strictly convex, continuously differentiable function that achieves its minimum value over $\mathbb{R}$ at some point in $(\underline{V}, \bar{V})$. It follows from properties (7.10) and (7.11) that $J(\cdot)$ has a unique minimizer v over $\mathbb{R}$ and it coincides with that for $\tilde{J}(\cdot)$. This minimizer v is the unique solution of the equation

$$\nabla \tilde{J}(v) = 0. \tag{7.12}$$

We calculate

$$\nabla \tilde{J}(w) = m_2K - m_1 + \frac{m_1^2w}{(m_1L + \frac{1}{2}m_1^2w^2)^{1/2}} + \frac{(Km_2)^2w}{(m_2L + \frac{1}{2}m_2^2K^2w^2)^{1/2}}.$$

Thus the optimal control policy for the original problem is that for problem $E_5(v)$, where v satisfies (7.12).

It is interesting to note that condition (7.1) in Proposition 7.1 determining the optimum velocity v at dropoff or pickup, which was previously derived from the optimal multiprocess maximum principle, can be interpreted as a statement that the gradient of the function $\tilde{J}$ (related to the value function) vanishes at v.

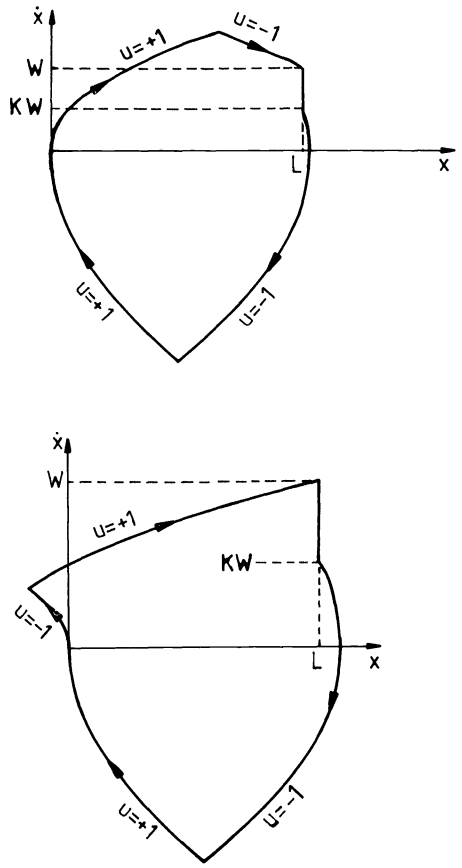

FIG. 2. *Robot arm problem: optimal phase plane trajectory when w is fixed (two switches).* $((2L/K^2m_1)^{1/2} \geqq w \geqq -(2L/K^2m_2)^{1/2}$, $m_2 = \frac{1}{2}m_1$, $K = \frac{1}{2})$. (b) *Robot arm problem: optimal phase plane trajectory when w is fixed (three switches).* $(w > (2L/K^2m_1)^{1/2}$, $m_2 = \frac{1}{2}m_1$, $K = \frac{1}{2})$.

Of course our success in solving problem (E_5) by traditional "two stage" optimization techniques very much depended on availability of a simple explicit formula for the value function. (It would be difficult to use these techniques when we refined the model, say to include damping and higher-order dynamics.) By contrast the theory of optimal multiprocesses would provide conditions on the solutions, even in the absence of such a formula.

Another point in favour of the theory of optimal multiprocesses as it bears on problem (E_5), is that it gives us certain information, prior to detailed analysis, about the number of switches in the optimal control policy (there are at most two of them). The two stage approach required us to consider the possibility of three switches, if only later to reject it, with consequent extra burden of analysis.

8. A problem in renewable resources. The most widely used tool in the theory of renewable resources is the Gordon-Schaefer model and its many variations (see, e.g., [3], [5]). Before giving the variation which constitutes our present example, we recapitulate the basic model. A population whose size x varies over time evolves

according to the law

$$\dot{x}(t) = F(x(t)) - \sigma x(t)u(t),$$

where F is a given (natural growth) function, u represents harvesting effort by the exploiter of the resource, and σ is a positive constant. Effort u is constrained to a given interval $[0, E]$. The initial value $x(0) = x_0$ is known. The exploiter's net return over a given interval of time $[0, T]$ is given in present value terms by

$$\int_0^T e^{-\delta t}\{\pi x(t) - c\}u(t)\, dt,$$

where δ is the discount rate, π the unit resource price, and c the unit effort cost. The problem is to choose the effort profile $u(\cdot)$ to maximize this quantity.

Under standard hypotheses that we omit, the solution to the problem is known to be of "turnpike" type. Specifically, there is a certain population level x^* (the formula for which we omit) in terms of which the solution may be described as follows (for definiteness, let us suppose that x_0 exceeds x^*). Initially maximum effort is applied ($u = E$) until $x(t)$ is driven down to x^*, at time $t = s_1$ (say). Then $x(t)$ is kept at the value x^* (by the appropriate effort) until $t = s_2$. And between s_2 and T there is another period of maximum effort ($u = E$). If the horizon T is too short, the intermediate (turnpike) portion of the solution may be absent.

The switching times s_1, s_2 are determined implicitly by certain boundary-value problems associated to the data of the model (in fact, s_2 is characterized in terms of the adjoint equation of the maximum principle together with the standard transversality condition).

Suppose now that a further possibility is open to the exploiter, that of shifting the operation from the initial population (which we shall now designate x_1, with data x_0^1, F_1, σ_1, π_1, c_1) to a different population x_2 (with data x_0^2, F_2, σ_2, π_2, c_2). The shift, which can occur at any chosen time τ, and that will be assumed instantaneous for simplicity, entails a lump cost ϕ_0 (at time τ). The problem then becomes that of minimizing

$$\phi_0\, e^{-\delta\tau} - \int_0^\tau e^{-\delta t}\{\pi_1 x_1 - c_1\}u\, dt - \int_\tau^T e^{-\delta t}\{\pi_2 x_2 - c_2\}u\, dt$$

over the admissible controls $u(\cdot)$ on $[0, T]$ and shift times τ in $[0, T]$. The dynamics are the initial ones on $[0, \tau)$, and those corresponding to the second population on $[\tau, T]$. Note that the initial condition appropriate to the dynamics on $[\tau, T]$ is given by $x_2(\tau) = z(\tau)$, where $z(\cdot)$ is the solution to

$$\dot{z}(t) = F_2(z(t)), \qquad z(0) = x_0^2. \tag{8.1}$$

We shall proceed to analyse the problem, assuming for simplicity that x_0^1 and x_0^2 are relatively large (i.e., exceed x_1^* and x_2^*, respectively).

It is clear that the problem is one of multiprocesses, with $k = 2$, $n_1 = n_2 = m_1 = m_2 = 1$. The set Λ and function f are given by

$$\{[0, \tau, x_0^1, x, \tau, T, z(\tau), y]\colon \tau \in [0, T],\ x \in \mathbb{R},\ y \in \mathbb{R}\},$$

$$f(\tau_0^1, \tau_1^1, x_0^1, x_1^1, \tau_0^2, \tau_1^2, x_0^2, x_1^2) = \phi_0\, e^{-\delta\tau_1^1}.$$

We suppose that the solution to the problem incorporates a shift at τ in $(0, T)$. At any point (τ, τ, z) in the set $\{[\tau, \tau, z(\tau)]\colon 0 \leqq \tau \leqq T\}$ having $0 < \tau < T$, the normal cone (space)

is spanned by the vector $[1, -1, 0]$ and $[1, 0, -\dot{z}(\tau)]$. With this in mind, it can be seen that the transversality condition of Theorem 3.1 yields in the present case:

$$h_0^2 = h_1^1 + p_2(\tau)\dot{z}(\tau) + \delta\phi_0\, e^{-\delta t}, \tag{8.2}$$

$$p_1(\tau) = p_2(T) = 0. \tag{8.3}$$

The latter conclusion (8.3) is familiar and expected from the original single-process problem, and expresses the fact that the imputed shadow price p is zero at the end of the planning period. This remains so for p_1 and p_2 because of the fact that (x_1, u) is optimal for the original problem restricted to $[0, \tau]$, while (x_2, u) is optimal for the second population studied on $[\tau, T]$.

Let us now examine (8.2), which constitutes the essential new information. From the single-process case it is known that (regardless of the lengths of the intervals $[0, \tau]$ and $[\tau, T]$) $u(t) = E$ on some interval to the left of τ and also on some interval to the right of τ. Thus we have

$$h_0^2 = p_2(\tau)[F_2(z(\tau)) - \sigma_2 z(\tau)E] + e^{-\delta\tau}[\pi_2 z(\tau) - c_2]E, \qquad h_1^1 = e^{-\delta\tau}[\pi_1 x_1(\tau) - c_1]E$$

(where we have used $p_1(\tau) = 0$). Substituting these expressions into (8.2) along with $\dot{z}(\tau) = F_2(z(\tau))$, we arrive at

$$[\pi_2 x_2(\tau) - c_2]E = [\pi_1 x_1(\tau) - c_1]E + \delta\phi_0 + e^{\delta\tau} p_2(\tau)\sigma_2 x_2(\tau)E. \tag{8.4}$$

This condition, which determines the shift time τ, can be given an economic interpretation. The (rate of) marginal revenue at τ from harvesting x_1 is given by $[\pi_1 x_1 - c_1]E$. The condition above states that when the shift occurs, the corresponding value for x_2, i.e., $[\pi_2 x_2 - c_2]E$, must equal that for x_1 plus a term corresponding to the (marginal) cost of making the shift (i.e., $\delta\phi_0$), plus a term $(e^{\delta\tau} p_2(\tau)\sigma_2 x_2(\tau)E)$ which is the (current) shadow value of the increased potential revenue from x_2 if x_2 were continued to be left to grow instead of being harvested.

Just as in the original single-process problem, the solution is completely but implicitly defined by the necessary conditions. In fact, much the same ingredients are involved in (8.4). For example, $x_1(\tau)$ is the final value of the optimal state for the single-process problem on $[0, \tau]$. This is known to coincide with $\max[y(\tau), x_f]$, where $y(\cdot)$ satisfies

$$\dot{y}(t) = F_1(y(t)) - \sigma_1 E y(t), \qquad y(0) = x_0^1$$

and where x_f is a certain constant determined by the data. Thus $x_1(\tau)$ is accessible, as is the adjoint value $p_2(\tau)$, although in this case it is a certain two-dimensional state-adjoint system that determines it. Finally, the value $x_2(\tau)$ in (8.4) is simply given by (8.1). In summary then, (8.4) permits us to determine the shift time τ. We remark that in this example it does not appear to be practical to solve the problem in a two-stage fashion, in view of the absence of explicit formulae for the appropriate value functions. On the other hand, the multiprocess necessary conditions have led directly to a characterization of the solution.

REFERENCES

[1] J.-P. AUBIN, *Mathematical Methods of Game and Economic Theory*, North-Holland, Amsterdam, 1979.
[2] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.
[3] C. W. CLARK, *Mathematical Bioeconomics*, John Wiley, New York, 1975.
[4] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.
[5] F. H. CLARKE AND G. R. MUNRO, *Coastal states, distant water fishing nations and extended jurisdiction: a principal-agent analysis*, Natural Resource Modelling, 2 (1987), pp. 81–107.

[6] F. H. CLARKE AND R. B. VINTER, *Optimal multiprocesses*, SIAM J. Control Optim., this issue, pp. 1072-1091.
[7] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[8] H. FRANKOWSKA, *Contingent cones to reachable sets of control systems*, preprint.
[9] R. V. GAMKRELIDZE, *On some extremal problems in the theory of differential equations with applications to the theory of optimal control*, SIAM J. Control, 3 (1965), pp. 106-128.
[10] R. GONZALES, *Sur la résolution de l'équation de Hamilton-Jacobi du contrôle deterministe*, Thèse, Université de Paris IX, 1980, Cahiers de Math de la Décision, Ceremade, 8029 bis.
[11] D. S. HAGUE, *Solution of multiple arc problems by the steepest descent method*, in Recent Advances in Optimization Techniques, A. Lavi and T. P. Vogl, eds., John Wiley, New York, 1965, pp. 489-518.
[12] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.
[13] D. M. KEIRSEY AND J. S. B. MITCHELL, *Planning strategic paths through variable terrain data*, in Proc. SPIE Applications of Artificial Intelligence No. 485, Arlington, VA, 1984, pp. 172-179.
[14] E. B. LEE AND L. MARCUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[15] D. Q. MAYNE AND E. J. POLAK, *A first order strong variation algorithm for optimal control*, J. Optim. Theory Appl., 16 (1975), pp. 277-301.
[16] J. S. B. MITCHELL, D. M. MOUNT, AND C. H. PAPADIMITRIOU, *The discrete geodesic problem*, SIAM J. Comput., 16 (1987), pp. 647-668.
[17] L. W. NEUSTADT, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.
[18] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Multiprocesses* (translated by K. N. Trirogoff; L. W. Neustadt, ed.), John Wiley, New York, 1962.
[19] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, SIAM J. Control Optim., 3 (1965), pp. 191-205.
[20] R. T. ROCKAFELLAR, *Optimality conditions for convex control problems with nonnegative states and the possibility of jumps*, in Game Theory and Mathematical Economics, O. Moeschlin and D. Pallaschke, eds., North-Holland, Amsterdam, 1981, pp. 339-349.
[21] K. TOMIYAMA, *Two-stage optimal control problems and optimality conditions*, J. Econom. Dynamics Control, 9 (1985), pp. 317-338.
[22] K. TOMIYAMA, *Multiple-stage optimal control problems with applications to plasma heating by neutron injection*, Technical Report UCLA-ENG-7780, School of Engineering and Applied Science, University of California, Los Angeles, CA, 1977.
[23] R. B. VINTER AND F. M. F. L. PEREIRA, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205-229.
[24] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[25] ———, *Steepest descent with relaxed controls*, SIAM J. Control Optim., 15 (1977), pp. 674-682.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1072-1091, September 1989

010

OPTIMAL MULTIPROCESSES*

FRANK H. CLARKE† AND RICHARD B. VINTER‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. A theory of necessary conditions for optimal multiprocesses is presented. Optimal multiprocesses are solutions to dynamic optimization problems described by families of control systems coupled through the boundary conditions and cost functions. The theory treats in a unified fashion a wide range of nonstandard dynamic optimization problems, and in many cases provides new optimality conditions. These include problems arising in impulse control, robotics, and optimal investment. Even when specialized to the (single process) free time optimal control problem, the theory improves on known necessary conditions. Detailed analysis of a number of applications and special cases appears in a companion paper.

Key words. optimal control, necessary conditions, differential inclusions

AMS(MOS) subject classifications. 49B10, 49B99

1. Introduction.

The theory presented in this paper originates in the authors' efforts in recent years to solve a variety of nonstandard problems in dynamic optimization. It became increasingly apparent in this earlier work that, although the problems considered were at first sight distinct, the task of finding good first-order conditions of optimality in each case called for development of similar analytical techniques. It was natural then to study these problems in a single comprehensive framework. We refer to this framework as multiprocesses.

Our companion paper [5], in which we analyse in detail a number of specific optimal multiprocess problems and special cases, testifies to the scope of the theory. The common feature of these problems is a family of dynamical equations with control term, coupled by a constraint on the boundary values of the constituent state trajectories and also by a function of these boundary values in the cost. We briefly describe some examples.

Consider the problem of determining an optimal control strategy for a multistage rocket, when the stage ejection times are included in the choice variables. (See, e.g., [8].) Standard optimality conditions of control theory, which accommodate changes in the dynamics only at fixed times, cannot be applied directly to such problems. It is true that we can sometimes overcome these difficulties by means of ad hoc techniques where, to begin with, we solve a family of problems involving fixed ejection times and then minimize over the ejection times ("two stage" optimization) or, alternatively, where we reduce the free ejection times to fixed ones by transformation of the time variable. To carry out two-stage optimization however we require tractable formulae describing the value function relative to the ejection times; unfortunately we cannot expect to derive such formulae for complicated problems. As for time transformation techniques, these are typically applicable only when certain regularity conditions are imposed on the data and this rules out certain significant applications. But if we view trajectory segments between ejection times as component trajectories, we arrive at an optimal multiprocess problem, for which our new theory supplies optimality conditions under very mild hypotheses.

* Received by the editors November 16, 1987; accepted for publication (in revised form)August 23, 1988.
† Centre de Recherches Mathématiques, Université de Montréal, Montréal, Québec, Canada H3C 3J7.
‡ Department of Electrical Engineering, Imperial College, London SW7 2BT, United Kingdom.

Next we consider a modification of the standard optimal control problem in which state trajectories are permitted to be discontinuous at a finite number of times, and the jump times are choice variables. The jump times, as well as the end states of the jumps, appear in the cost and are constrained to lie in some closed set. There is also a constraint on the number of jumps. Problems such as this arise in optimal investment and inventory control, and have been called "impulse control" problems. Necessary conditions of optimality are available (see, e.g., [11]-[15]) but these are limited to only a narrow class of cost functions and constraints, as far as the jumps are concerned, and apply only when the constraint on the number of jumps is inactive. We can view these impulse control problems as another instance of an optimal multiprocess problem, however. Here we interpret restrictions of trajectories (in the original problem) to intervals between jumps as component trajectories (in the optimal multiprocess problem). We thereby obtain necessary conditions of optimality when restrictions in the earlier theory, on the manner in which the jumps are included in the cost and constraints, are lifted.

An example of a variational problem to which our theory is applicable is the derivation of Snell's law of refraction from Fermat's principle of least time. If the two media involved are homogeneous we derive a simple formula for the minimum time as a function of the point of incidence, and deduce Snell's law by setting the gradient to zero. For inhomogeneous media this "two stage" approach is no longer available to us outside the simplest cases. However we can proceed alternatively by regarding the light ray as solving an optimal multiprocess problem: the component trajectories are the path segments in each medium and the constraint on the boundary values is that the path segments match at the interface. We can thereby validate Snell's Law in a very general setting.

It is remarkable that, even when we specialize the theory to treatment of the standard optimal control problem, significant advances on known results ensue. For free time optimal control problems the Pontryagin maximum principle provides a boundary condition on the maximized Hamiltonian, evaluated along the optimal trajectory. The boundary condition has previously been proved only under the assumption that the data is continuous in the time variable. Multiprocess theory, by contrast, yields a version of the boundary condition when the data is merely measurable in the time variable.

There is a long tradition of research whose aim is unification of optimality conditions for as wide a range of optimization problems as possible (in mathematical programming, control, and the calculus of variations), within a single general theory. Notable contributions here are Neustadt's abstract variational theory [10] and the general multiplier rules of Warga [14] and Ioffe [9]. The price paid for the comprehensiveness of these theories is the labour involved in checking hypotheses (most notably in applications to optimal control) and, in the case of Neustadt's theory, the ingenuity required in devising certain approximating sets. The theory of optimal multiprocesses is a general theory and so must be considered as partly in this tradition. But it is a departure too, since it is targeted specifically at problems in dynamic optimization. Consequently, for such problems, checking hypotheses is usually a trivial task. Our applications in [5] illustrate that generating optimality conditions for specific classes of problems is usually a routine matter of calculating normal cones.

We make some comments concerning the analytical techniques employed in this paper. These follow the lead of a number of recent publications in which optimality conditions are proved by considering limits of normal vectors to the epigraph of a suitable value function. This approach ("proximal normal analysis"), which was first

adopted in [2], besides being useful for proving necessary conditions, provides new interpretations of the Lagrange multipliers involved in terms of sensitivity of the minimum cost to data perturbations. Interpretations of this kind have been obtained by other methods in mathematical programming (see, e.g., [7], [12]) and via proximal normal analysis by Clarke and Loewen [3] for optimal control problems. The proofs of this paper illustrate the power of proximal normal analysis to also generate new optimality conditions.

Our results admit extensions in various directions. These will be the subject of future work. In particular, a similar but more intricate analysis leads to a maximum principle when unilateral state constraints are introduced in the optimal multiprocess problem. Under more stringent assumptions we can also provide an accompanying sensitivity analysis.

Finally we mention that a simple proof can be given of a restricted form of the optimal multiprocess maximum principle when the data is smooth in time. This is based on a reduction of the optimal multiprocess problem to a standard optimal control problem by transformation of the time variable. The restricted form of the principle is inadequate, however, for certain applications (see [5]).

2. Essential values. For free time dynamic optimization problems, such as those studied here, we can expect optimality conditions to incorporate conditions on boundary values of functions constructed from the costate variables. When the data is merely measurable in the time variable, the elementary definition of boundary value, namely evaluation at a point, cannot be adopted, basically because it is not robust under the limiting arguments employed in derivation of optimality conditions. Instead we must interpret boundary values as "essential values."

Let $S \subset \mathbb{R}$ be an open set, T a point in S, and $\psi : S \to \mathbb{R}^k$ a measurable function. The set of essential values of ψ at T, denoted $\operatorname{ess}_{t \to T} \psi(t)$, is defined as follows. ζ belongs to this set if and only if, for any positive number $\varepsilon > 0$, the following set has positive Lebesgue measure:

$$\{t: T - \varepsilon < t < T + \varepsilon, |\zeta - \psi(t)| < \varepsilon\}.$$

If a point lies in $\operatorname{co} \operatorname{ess}_{t \to T} \psi(t)$ we say it is a convex essential value of ψ at T.

It is clearly the case that if ψ is continuous at T then

$$\operatorname*{ess}_{t \to T} \psi(t) = \{\psi(T)\},$$

i.e., the essential value is merely the value of the function.

Given a set $D \subset \mathbb{R}^l$ and a multifunction $A : D \rightrightarrows \mathbb{R}^k$, we say that A is *closed* if, for any convergent subsequences $\{y_i\} \subset D$ and $\{a_i\} \subset \mathbb{R}^k$ (we write the limits y and a, respectively) such that $a_i \in Ay_i$ for $i = 1, 2, \cdots$ and $y \in D$, we have $a \in Ay$. The following closedness property of the multifunction obtained by applying the operation of taking essential values of a function accounts for its significance in optimization theory.

LEMMA 2.1. *Let P, Q be open subsets of $\mathbb{R}$, $\mathbb{R}^n$, respectively, and let $h : P \times Q \to \mathbb{R}^k$ be a given function. Suppose*

$$x \to h(t, x) \text{ is continuous, uniformly in } t, \text{ and}$$

$$t \to h(t, x) \text{ is measurable for every } x \in Q.$$

Then the multifunction $G : P \times Q \rightrightarrows \mathbb{R}^k$ defined by $G(t, x) = \operatorname{ess}_{s \to t} h(s, x)$ is closed. If in addition we have

$$\sup_{x \in P} \operatorname*{ess}_{s \to t} |h(s, x)| < \infty,$$

then $(t, x) \to \operatorname{co} G(t, x)$ is also a closed multifunction. Here "co" denotes convex hull.

Proof. Consider the first assertion. Let $t_i \to t$, $x_i \to x$ and $r_i \to r$ where $t \in P$, $x \in Q$, and $r_i \in \operatorname{ess}_{s \to t_i} h(s, x_i)$ for each i, $t \in P$, and $x \in Q$. We must show that $r \in \operatorname{ess}_{s \to t} h(s, x)$. Choose $\varepsilon > 0$ and define

$$S_i^\varepsilon = \{s \in (t_i - \varepsilon/2,\ t_i + \varepsilon/2) \cap P \,|\, |h(s, x_i) - r_i| < \varepsilon/2\}.$$

By definition of essential values, the set S_i^ε has positive measure. But for i sufficiently large $|t_i - t| < \varepsilon/2$ and $|h(s, x_i) - h(s, x)| + |r - r_i| < \varepsilon/2$ for all $s \in P$ by uniform continuity. It follows that

$$S_i^\varepsilon \subset S^\varepsilon$$

where

$$S^\varepsilon = \{s \in (t - \varepsilon,\ t + \varepsilon) \cap P \,|\, |h(s, x) - r| < \varepsilon\}.$$

The set S^ε then has positive measure. Since ε is arbitrary, $r \in \operatorname{ess}_{s \to t} h(s, t)$. A simple compactness argument, and application of Caratheodory's Theorem, now give the second assertion.

3. A maximum principle for optimal multiprocesses. Frequent reference is made to points in product spaces, and to products of product spaces. In this connection, a point $((a_1, b_1, \cdots), (a_2, b_2, \cdots), \cdots, (a_k, b_k, \cdots))$ is denoted by $\{a_i, b_i, \cdots\}_{i=1}^k$ or, briefly, $\{a_i, b_i, \cdots\}$.

The following data are given:

positive integers k, and n_i, m_i,	$i = 1, \cdots, k$,
functions $\phi_i : \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{m_i} \to \mathbb{R}^{n_i}$,	$i = 1, \cdots, k$,
subsets U^i of $\mathbb{R} \times \mathbb{R}^{m_i}$,	$i = 1, \cdots, k$,
subsets X^i of $\mathbb{R} \times \mathbb{R}^{n_i}$,	$i = 1, \cdots, k$.

We term multiprocess a point $\{\tau_0^i, \tau_1^i, x_i(\cdot), w_i(\cdot)\}$ comprising left and right endpoints τ_0^i, τ_1^i of a closed subinterval of $\mathbb{R}$, absolutely continuous functions $x_i(\cdot) : [\tau_0^i, \tau_1^i] \to \mathbb{R}^{n_i}$ and measurable functions $w_i(\cdot) : [\tau_0^i, \tau_1^i] \to \mathbb{R}^{m_i}$ such that

$$\begin{aligned}
&\dot{x}_i(t) = \phi_i(t, x_i(t), w_i(t)) && \text{a.e. } t \in [\tau_0^i, \tau_1^i], \\
&w_i(t) \in U_t^i, && \text{a.e. } t \in [\tau_0^i, \tau_1^i], \\
&x_i(t) \in X_t^i, && \text{for all } t \in [\tau_0^i, \tau_1^i],
\end{aligned}$$

for $i = 1, \cdots, k$. Here U_t^i is the set $\{u \,|\, (t, u) \in U^i\}$, and X_t^i is likewise defined.

It is assumed that the data satisfies the following hypotheses.

(H1) For each $x \in \mathbb{R}^{n_i}$, $\phi_i(\cdot, x, \cdot)$ is $\mathscr{L} \times \mathscr{B}$ measurable (where $\mathscr{L}$ is the class of Lebesgue subsets of $\mathbb{R}$ and $\mathscr{B}$, the class of Borel subsets of $\mathbb{R}^{m_i}$) for $i = 1, \cdots, k$.

(H2) U^i is a Borel measurable set for $i = 1, \cdots, k$.

There exists a constant K with the following properties.

(H3) $|\phi_i(t, y, w)| \leqq K$ whenever $(t, y, w) \in \mathbb{R} \times X_t^i \times U_t^i$.

(H4) $|\phi_i(t, y, w) - \phi_i(t, y', w)| \leqq K |y - y'|$ whenever (t, y, w), $(t, y', w) \in \mathbb{R} \times X_t^i \times U_t^i$.

We aim to give optimality conditions for an optimization problem posed over multiprocesses. However, as is common, we approach the optimization problem via analysis of boundary points of a reachable set.

Let C be a given closed set in

$$\prod_i \{(\tau_0^i, \tau_1^i, a_0^i) \mid a_0^i \in \mathbb{R}^{n_i},\ \tau_0^i, \tau_1^i \in \mathbb{R},\ \tau_0^i \leqq \tau_1^i\}$$

and let $\psi : \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}^d$ be a given Lipschitz continuous function. We define the reachable set (with respect to C and ψ), written $\mathcal{R}_{\psi,C}$, to be

$$\mathcal{R}_{\psi,C} := \{\psi(\{y_i(\tau_1^i)\}) \mid \{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\} \text{ is a multiprocess such that } \{\tau_0^i, \tau_1^i, y_i(\tau_0^i)\} \in C\}.$$

We say that a multiprocess $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ is a boundary multiprocess relative to ψ and C if

$$\{\tau_0^i, \tau_1^i, y_i(\tau_0^i)\} \in C \quad \text{and} \quad \psi(\{y_i(\tau_1^i)\}) \in \partial \mathcal{R}_{\psi,C}$$

(∂ denotes boundary). Define the unmaximized Hamiltonian function H_i to be

$$H_i(t, x, u, p) := p \cdot \phi_i(t, x, u), \qquad i = 1, \cdots, k.$$

The following theorem is a necessary condition that a multiprocess be associated with a boundary point of the reachable set.

THEOREM 3.1. *Let $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ be a boundary multiprocess (with respect to C and ψ). Assume that*

$$\text{graph}\,\{x_i(\cdot)\} \subset \text{interior}\,\{X^i\}$$

for $i = 1, \cdots, k$ and that hypotheses (H1)-(H4) *are satisfied. Then there exists a vector v of unit length, numbers h_0^i, h_1^i and absolutely continuous functions $p_i(\cdot) : [T_0^i, T_1^i] \to \mathbb{R}^{n_i}$ for $i = 1, \cdots, k$, and a number c (whose magnitude is governed by the constant K in hypotheses* (H3) *and* (H4) *together with the Lipschitz rank of ψ restricted to some neighbourhood of $\{x_i(T_1^i)\}$), with the following properties*:

$$-\dot{p}_i(t) \in \partial_x H_i(t, x_i(t), u_i(t), p_i(t)) \quad \text{a.e. } t \in [T_0^i, T_1^i],$$

$$H_i(t, x_i(t), u_i(t), p_i(t)) = \max_{w \in U_t^i} H_i(t, x_i(t), w, p_i(t)) \quad \text{a.e. } t \in [T_0^i, T_1^i],$$

$$h_0^i \in \text{co ess}_{t \to T_0^i} \left[\sup_{w \in U_t^i} H_i(t, x_i(T_0^i), w, p_i(T_0^i)) \right],$$

$$h_1^i \in \text{co ess}_{t \to T_1^i} \left[\sup_{w \in U_t^i} H_i(t, x_i(T_1^i), w, p_i(T_1^i)) \right]$$

for $i = 1, \cdots, k$,

$$\{p_i(T_1^i)\} \in \partial \psi^*(\{x_i(T_1^i)\}) v$$

and

$$\{-h_0^i, h_1^i, p_i(T_0^i)\} \in c \partial d_C(\{T_0^i, T_1^i, x_i(T_0^i)\}).$$

Here $\partial_x H_i$ denotes the partial generalized gradient in the second variable and $\partial \psi^*$ is the transpose of the generalized Jacobian of ψ. (Generalized gradients and Jacobians are understood in the sense of Clarke [2]). *The operation of taking essential values* "ess" *has been defined in* § 2. *d_C is the Euclidean distance function from the set C.*

The theorem is proved in § 6.

Let

$$f: \quad \prod_i (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{n_i}) \to \mathbb{R}$$

be a given locally Lipschitz continuous function and let

$$\Lambda \subset \prod_i \{(\tau_0^i, \tau_1^i, a_0^i, a_1^i) | \tau_0^i, \tau_1^i \in \mathbb{R}, a_0^i, a_1^i \in \mathbb{R}^{n_i}, \tau_0^i \leqq \tau_1^i\}$$

be a given closed set.

We now pose the optimal multiprocess problem:

$$(P) \quad \begin{array}{l} \text{minimize } f(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\}) \text{ over multiprocesses } \{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\} \\ \text{that satisfy } \{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\} \subset \Lambda. \end{array}$$

In § 7 we shall derive from Theorem 3.1 the following maximum principle for solutions to the optimal multiprocess problem.

THEOREM 3.2. *Let* $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ *be a solution to* (P). *Assume that*

$$\text{graph } \{x_i(\cdot)\} \subset \text{interior } \{X^i\}$$

for $i = 1, \cdots, k$ *and that hypotheses* (H1)-(H4) *are satisfied. Then there exists a real number* $\lambda \geqq 0$, *real numbers* h_0^i, h_1^i, *and absolutely continuous functions* $p_i(\cdot) : [T_0^i, T_1^i] \to \mathbb{R}^{n_i}$ *for* $i = 1, \cdots, k$, *and a constant* c (*whose magnitude is determined by the constant* K *of hypotheses* (H3) *and* (H4) *together with the Lipschitz rank of* f *in a neighbourhood of* $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$) *such that* $\lambda + \sum_i |p_i(T_1^i)| = 1$ *and we have*

$$(3.1) \quad -\dot{p}_i(t) \in \partial_x H_i(t, x_i(t), u_i(t), p_i(t)) \quad \text{a.e. } t \in [T_0^i, T_1^i],$$

$$H_i(t, x_i(t), u_i(t), p_i(t)) = \max_{w \in U_t^i} H_i(t, x_i(t), w, p_i(t)) \quad \text{a.e. } t \in [T_0^i, T_1^i],$$

$$(3.2) \quad h_0^i \in \text{co ess}_{t \to T_0^i} \left[\sup_{w \in U_t^i} H_i(t, x_i(T_0^i), w, p_i(T_0^i)) \right],$$

$$h_1^i \in \text{co ess}_{t \to T_1^i} \left[\sup_{w \in U_t^i} H_i(t, x_i(T_1^i), w, p_i(T_1^i)) \right]$$

for $i = 1, \cdots, k$, *and*

$$(3.3) \quad \{-h_0^i, h_1^i, p_i(T_0^i), -p_i(T_1^i)\} \in c\, \partial d_\Lambda + \lambda \partial f$$

where the generalized gradients ∂d_Λ *and* ∂f *are evaluated at* $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$.

Note that all the ingredients of the traditional maximum principle, namely costate functions $p(\cdot)$, costate differential inclusions (3.1), and maximization of the Hamiltonian (3.2), are present in the multiprocess maximum principle. The costate differential inclusions and the Hamiltonian maximizing properties separate out into statements about the individual component processes. The fact that the component processes in the optimal multiprocess problem are coupled through an endpoint constraint and the cost function gives rise to a corresponding coupling of the component costate functions, the $p_i(\cdot)$'s, through their endpoints via the multiprocess transversality condition (3.3).

Our formulation of the optimization problem (P) incorporates the constraints "$y_i(t) \in X_t^i$" mainly for the purpose of hypothesis refinement. A consequence of our theorem is that the assertions remain valid if $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ is merely a local solution to (P), in the sense that it is minimizing with respect to all multiprocesses $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ that satisfy

$$\text{graph } \{y_i(\cdot)\} \subset \text{graph } \{x_i(\cdot)\} + \varepsilon B$$

for some $\varepsilon > 0$ (B denotes the open unit ball). It follows also that the boundedness and uniform Lipschitz continuity hypotheses ((H3) and (H4)) need to be checked only on neighbourhoods of the graphs of the minimizing $x_i(\cdot)$'s. (For confirmation we have

only to replace each X_i by its intersection with the relevant neighbourhood of graph $\{x_i\}$ and apply the original theorem.) In the event that $x_i(\cdot)$ strikes the boundary of the set X_i for some i, the hypotheses of Theorem 3.2 are violated; in such circumstances necessary conditions of optimality, which involve possible discontinuous component costate functions $\{p_i(\cdot)\}$, may be derived. We do not pursue such extensions here.

4. Coupled dynamic optimization problems: a differential inclusion formulation. It is well known that we may choose a variety of starting points for derivation of conditions on solutions to dynamic optimization problems over a single time interval. Two notable instances are, first, that where the dynamics are modeled by a differential equation with control and, second, that involving a differential inclusion. The first-order optimality conditions derivable for each of these formulations are distinct; examples of problems are known where the differential equations conditions give more information about solutions than the differential inclusions conditions, and vice versa.

Equally, distinct sets of necessary conditions for solutions to coupled dynamic optimization problems over a family of time intervals result, according to whether the dynamics are described by differential equations or differential inclusions. We have already given necessary conditions in the differential equations case. In this section we do the same for differential inclusions.

Our necessary conditions on solutions to coupled dynamic optimization problems in a differential inclusions context, in addition to providing independent information about solutions in certain cases, will be important here as constituting the first step in our proof of the maximum principle governing optimal multiprocesses.

The following data are given:

positive integers k, and n_i, $\quad i = 1, \cdots, k$,

a function $g: \prod_{i=1}^{k} (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{n_i}) \to \mathbb{R}$,

multifunctions $F_i : \mathbb{R} \times \mathbb{R}^{n_i} \rightrightarrows \mathbb{R}^{n_i}$, $\quad i = 1, \cdots, k$,

sets $\Gamma^i \subset \mathbb{R} \times \mathbb{R}^{n_i}$, $\quad i = 1, \cdots, k$,

and a subset M of

$$\prod_{i=1}^{k} \{(\tau_0^i, \tau_1^i, a_0^i, a_1^i) | \tau_0^i, \tau_1^i \in \mathbb{R}, a_0^i, a_1^i \in \mathbb{R}^{n_i} \text{ and } \tau_0^i \leqq \tau_1^i\}.$$

Now consider the following problem:

$$\begin{aligned} &\text{minimize} \quad g(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\}) \\ &\text{subject to} \\ (\mathrm{Q}) \qquad &\dot{y}_i(t) \in F_i(t, y_i(t)) \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i], \\ &\dot{y}_i(t) \in \Gamma_t^i \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i], \\ &\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\} \in M. \end{aligned}$$

The underlying elements in the minimization problem here are k-tuples $\{\tau_0^i, \tau_1^i, y_i(\cdot)\}_{i=1}^k$ in which τ_0^i and τ_1^i are, respectively, left and right endpoints of a closed subinterval of $\mathbb{R}$, and $y_i(\cdot)$ is an absolutely continuous n_i vector valued function on $[\tau_0^i, \tau_1^i]$, $i = 1, \cdots, k$.

The hypotheses invoked will be as follows:

(I1) g is locally Lipschitz continuous.

(I2) M is closed.

(I3) For each i, F_i takes values closed, convex sets, and given any point $x \in \mathbb{R}^{n_i}$ and closed set $D \subset \mathbb{R}^{n_i}$, the set $\{t \mid D \cap F_i(t, x) \neq \emptyset\}$ is Lebesgue measurable.

There exists a constant K such that we have the following:

(I4) $|v| \leqq K$ whenever $v \in F_i(t, x)$, $(t, x) \in \Gamma^i$, $i = 1, \cdots, k$.

(I5) $\operatorname{dist}\{F_i(t, x), F_i(t, y)\} \leqq K|x - y|$, whenever (t, x), $(t, y) \in \Gamma^i$, $i = 1, \cdots, k$ ("dist" is the Hausdorff distance function).

We define the Hamiltonian functions $\mathscr{H}_i : \Gamma^i \times \mathbb{R}^{n_i} \to \mathbb{R}$ to be

$$\mathscr{H}_i(t, x, p) := \sup_{e \in F_i(t,x)} p \cdot e, \qquad i = 1, \cdots, k.$$

THEOREM 4.1. *Let* $\{T_0^i, T_1^i, x_i(\cdot)\}$ *solve problem* (Q). *Assume that*

$$\operatorname{graph}\{x_i(\cdot)\} \subset \operatorname{interior}\{\Gamma^i\}$$

for $i = 1, \cdots, k$, *and that hypotheses* (I1)-(I5) *are satisfied. Then there exists a real number* $\lambda \geqq 0$, *real numbers* h_0^i, h_1^i, *absolutely continuous functions* $p_i(\cdot) : [T_0^i, T_1^i] \to \mathbb{R}^{n_i}$, $i = 1, \cdots, k$, *and a constant* c (*whose magnitude is determined by the constant* K *of hypotheses* (I4) *and* (I5) *together with the Lipschitz rank of* g *in a neighbourhood of* $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$), *such that* $\lambda + \sum_i |p_i(T_1^i)| = 1$ *and we have*

$$(-\dot{p}_i(t), \dot{x}_i(t)) \in \partial_{x,p} \mathscr{H}_i(t, x_i(t), p_i(t)) \quad \text{a.e. } t \in [T_0^i, T_1^i],$$

$$h_0^i \in \operatorname{co} \operatorname*{ess}_{t \to T_0^i} \mathscr{H}_i(t, x_i(T_0^i), p_i(T_0^i)), \; h_1^i \in \operatorname{co} \operatorname*{ess}_{t \to T_1^i} \mathscr{H}_i(t, x_i(T_1^i), p_i(T_1^i))$$

for $i = 1, \cdots, k$, *and* $\{-h_0^i, h_1^i, p_i(T_0^i), -p_i(T_1^i)\} \in c\, \partial d_M + \lambda\, \partial g$. *Here* $\partial_{x,p} \mathscr{H}_i$ *denotes the partial generalized gradient of* $\mathscr{H}_i$ *in the second and third variables. The generalized gradients* ∂d_C *and* ∂g *are evaluated at* $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$.

Theorem 4.1 is proved in § 5.

5. Proof of Theorem 4.1.

A special case. Our strategy is first to prove the theorem in the presence of two supplementary hypotheses. These will be disposed of at a later stage. For the time being then we impose the following hypotheses.

(IU) $\{(T_0^i, T_1^i, x_i(\cdot)\}$ is the unique solution to (Q).

(IL) g is a linear function of the form $g(\{\tau_0^i, \tau_1^i, y_0^i, y_1^i\}) = \sum_{i=1}^k g_i \cdot y_1^i$ in which g_i is a given vector in $\mathbb{R}^{n_i}$, $i = 1, \cdots, k$.

The proof hinges on an introduction of a family of problems $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$ generated by perturbations to the constraint set M. Choose $\varepsilon > 0$ (this will remain fixed) with the property

$$\operatorname{graph}\{x_i(\cdot)\} + 2\varepsilon B \subset \Gamma^i, \qquad i = 1, \cdots, k,$$

and define the closed sets $\tilde{\Gamma}^i$, $i = 1, \cdots, k$ to be

$$\tilde{\Gamma}^i := \operatorname{graph}\{x_i(\cdot)\} + \varepsilon \bar{B}.$$

For each vector $\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\} \in \prod_i (\mathbb{R} \times \mathbb{R} \times \mathbb{R}^{n_i} \times \mathbb{R}^{n_i})$ problem $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$ is taken to be the following:

$$\begin{aligned}
&\text{minimize} \quad g(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\}) \\
&\text{subject to} \\
&\qquad \dot{y}_i(t) \in F_i(t, y_i(t)) \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i], \\
&\qquad \dot{y}_i(t) = 0 \qquad\qquad\; \text{a.e. } t \in I_i \backslash [\tau_0^i, \tau_1^i], \\
&\qquad \operatorname{graph}\{y_i(\cdot)\} \subset \tilde{\Gamma}^i
\end{aligned}$$

for $i=1,\cdots,k$ and

$$\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\} \in M + \{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}. \tag{5.1}$$

In the above, set I_i is the fixed time interval

$$I_i = [T_0^i - \varepsilon,\ T_1^i + \varepsilon],$$

$i=1,\cdots,k$. Minimization is conducted over k-tuples of elements $\{\tau_0^i, \tau_1^i, y_i(\cdot)\}_{i=1}^k$ comprising Lipschitz continuous functions $y_i(\cdot): I_i \to \mathbb{R}^{n_i}$ and endpoints τ_0^i, τ_1^i of closed intervals satisfying $[\tau_0^i, \tau_1^i] \subset I_i$. A k-tuple of such elements satisfying the constraints of the problem, with the possible exception of (5.1), is called a *trajectory*. In the event (5.1) is satisfied as well, the trajectory is said to be *admissible* (for $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$. Problem $Q(\{0,0,0,0\})$ will be recognized as a refinement of problem (Q) in which the constraint set Γ^i is replaced by the closed subset $\tilde{\Gamma}^i$. Clearly the point $\{T_0^i, T_1^i, x_i(\cdot)\}$ remains a solution to $Q(\{0,0,0,0\})$.

We denote by V the value function associated with these perturbations of problem (Q): given $\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\} \in \prod_i (\mathbb{R}\times\mathbb{R}\times\mathbb{R}^{n_i}\times\mathbb{R}^{n_i})$ then $V(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$ is taken to be the infimum cost of problem $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$. (The infimum cost is interpreted as "$+\infty$" if no admissible trajectory exists.)

Standard sequential compactness arguments of existence theory (see, e.g., [2, Thm. 3.1.7]) together with (IU) yield the fact that $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$ has a solution if there is an admissible trajectory, along with the following information.

LEMMA 5.1. (i) *Let* $\{\bar{\rho}_0^i, \bar{\rho}_1^i, \bar{\sigma}_0^i, \bar{\sigma}_1^i\}$ *represent terms in a sequence converging to* $\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}$ *and let* $\{\bar{\tau}_0^i, \bar{\tau}_1^i, \bar{y}_i(\cdot)\}$ *be a solution to* $Q(\{\bar{\rho}_0^i, \bar{\rho}_1^i, \bar{\sigma}_0^i, \bar{\sigma}_1^i\})$. *Then, following replacement of the original sequence by a subsequence if necessary, we have that* $\bar{\tau}_0^i \to \tau_0^i$, $\bar{\tau}_1^i \to \tau_1^i$ *for each* i, *and* $\bar{y}_i(\cdot) \to y_i(\cdot)$ *uniformly for each* i *where* $\{\tau_0^i, \tau_1^i, y_i(\cdot)\}$ *is an admissible trajectory for* $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$.

(ii) *If in part* (i) $\{\rho_0^i, \rho_i^i, \sigma_0^i, \sigma_i^i\} = \{0,0,0,0\}$ *and also* $V(\{\bar{\rho}_0^i, \bar{\rho}_1^i, \bar{\sigma}_0^i, \bar{\sigma}_1^i\}) \to V\{0,0,0,0\}$ *then* $\{\tau_0^i, \tau_1^i, y_i(\cdot)\} = \{T_0^i, T_1^i, x_i(\cdot)\}$.

(iii) *The epigraph of* V, epi V, *is closed.*

We now proceed to an analysis of proximal normals to epi V at a point

$$[\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\},\ V(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}) + \delta]$$

(with $\delta \geqq 0$). Recall that a nonzero vector ζ is a proximal normal to a closed set $S \subset \mathbb{R}^q$ at one of its points s provided that, for some $m \geqq 0$, we have

$$-\zeta \cdot s' + m|s'-s|^2 \geqq -\zeta \cdot s \quad \text{for all } s' \in S. \tag{5.2}$$

In the present analysis the role of S is played by epi V and the only fact required about proximal normals is that a dense subset of points s in the boundary of S admit a proximal normal (or perpendicular [2, p. 66]).

LEMMA 5.2. *Let* $[\{h_0^i, -h_1^i, -s_0^i, s_1^i\}, -\lambda]$ *be a proximal normal to* epi V *at the point* $[\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\},\ V(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}) + \delta]$. *Let* $\{\tau_0^i, \tau_1^i, z_i(\cdot)\}$ *solve* $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$ *and suppose that graph* $\{z_i(\cdot): [\tau_0^i, \tau_1^i] \to \mathbb{R}^{n_i}\}$ *is interior to* $\tilde{\Gamma}^i$ *for* $i=1,\cdots,k$. *Let* $\{\alpha_0^i, \alpha_1^i, \gamma_0^i, \gamma_1^i\}$ *be the point in* M *such that* $\{\tau_0^i, \tau_1^i, z_i(\tau_0^i), z_i(\tau_1^i)\} = \{\alpha_0^i, \alpha_1^i, \gamma_0^i, \gamma_1^i\} + \{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}$. *Then for* $i=1,\cdots,k$ *there exists an absolutely continuous function* $p_i(\cdot): I_i \to \mathbb{R}^{n_i}$ *such that*

$$(-\dot{p}_i(t), \dot{z}_i(t)) \in \begin{cases} \partial \mathscr{H}_i(t, z_i(t), p_i(t)) & \text{a.e. on } [(\tau_0^i, \tau_1^i]), \\ \{0, 0\} & \text{a.e. } I_i \backslash [(\tau_0^i, \tau_1^i]), \end{cases} \tag{5.3}$$

$$p_i(\tau_0^i) = s_0^i, \tag{5.4}$$

$$p_i(\tau_1^i) = s_1^i - \lambda g_i, \tag{5.5}$$

$$h_0^i \in \operatorname{co} \operatorname*{ess}_{t \to \tau_0^i} \mathscr{H}_i(t, z_i(\tau_0^i), p_i(\tau_0^i)), \tag{5.6}$$

$$h_1^i \in \operatorname{co} \operatorname*{ess}_{t \to \tau_1^i} \mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)). \tag{5.7}$$

Furthermore,

$$\{h_0^i, -h_1^i, -s_0^i, s_1^i\} \in |\{h_0^i, -h_1^i, -s_0^i, s_1^i\}| \partial d_M(\{\alpha_0^i, \alpha_1^i, \gamma_0^i, \gamma_1^i\}). \tag{5.8}$$

Proof. Let $\{t_0^i, t_1^i, y_i(\cdot)\}$ be an arbitrary trajectory. Let $\{\bar{\alpha}_0^i, \bar{\alpha}_1^i, \bar{\gamma}_0^i, \bar{\gamma}_1^i\}$ be any point in M and $\bar{\delta}$ any nonnegative number. Observe that, by definition of the epigraph, the point

$$\left[\{t_0^i - \bar{\alpha}_0^i, t_1^i - \bar{\alpha}_1^i, y_i(t_0^i) - \bar{\gamma}_0^i, y_i(t_1^i) - \bar{\gamma}_1^i\}, \sum_i g_i \cdot y_i(t_1^i) + \bar{\delta}\right]$$

lies in epi V; we use this in the proximal normal inequality (5.2). The role of s is played by

$$\left[\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}, \sum_i g_i \cdot z_i(\tau_1^i) + \delta\right],$$

which equals

$$\left[\{\tau_0^i - \alpha_0^i, \tau_1^i - \alpha_1^i, z_i(\tau_0^i) - \gamma_0^i, z_i(\tau_1^i) - \gamma_1^i\}, \sum_i g_i z_i(\tau_1^i) + \delta\right]$$

and ζ is of course $[\{h_0^i, -h_1^i, -s_0^i, s_1^i\}, -\lambda]$.

Substitution into (5.2) leads to the following conclusions:

$$\begin{aligned} \sum_i \Big[& -h_0^i(t_0^i - \bar{\alpha}_0^i - \tau_0^i + \alpha_0^i) + h_1^i(t_1^i - \bar{\alpha}_1^i - \tau_1^i + \alpha_1^i) \\ & + s_0^i \cdot (y_i(t_0^i) - \bar{\gamma}_0^i - z_i(\tau_0^i) + \gamma_0^i) - s_1^i \cdot (y_i(t_1^i) - \bar{\gamma}_1^i - z_i(\tau_1^i) + \gamma_1^i) \\ & + \lambda \left(\sum_i g_i \cdot y_i(t_1^i) + \bar{\delta} - \sum_i g_i \cdot z_i(\tau_1^i) - \delta \right) \Big] + m\Delta \geqq 0. \end{aligned} \tag{5.9}$$

Here m is some nonnegative number and

$$\begin{aligned} \Delta = & \left| \sum_i g_i \cdot y_i(t_1^i) + \bar{\delta} - \sum_i g_i \cdot z_i(\tau_1^i) - \delta \right|^2 \\ & + \sum_i (|t_0^i - \bar{\alpha}_0^i - \tau_0^i + \alpha_0^i|^2 + |t_1^i - \bar{\alpha}_1^i - \bar{\tau}_1^i + \alpha_1^i|^2) \\ & + \sum_i (|y_i(t_0^i) - \bar{\gamma}_0^i - z_i(\tau_0^i) + \gamma_0^i|^2 + |y_i(t_1^i) - \bar{\gamma}_1^i - z_i(\tau_1^i) + \gamma_1^i|^2). \end{aligned}$$

In (5.9) we set

$$\{t_0^i, t_1^i, y_i(\cdot)\} = \{\tau_0^i, \tau_1^i, z_i(\cdot)\}$$

to derive

$$\sum_i (-h_0^i(\alpha_0^i - \bar{\alpha}_0^i) + h_1^i(\alpha_1^i - \bar{\alpha}_1^i) + s_0^i(\gamma_0^i - \bar{\gamma}_0^i) - s_1^i(\gamma_1^i - \bar{\gamma}_1^i)) + \lambda(\bar{\delta} - \delta)$$

$$+ m\left(\sum_i (|\alpha_0^i - \bar{\alpha}_0^i|^2 + |\alpha_1^i - \bar{\alpha}_1^i|^2 + |\gamma_0^i - \bar{\gamma}_0^i|^2 + |\gamma_1^i - \bar{\gamma}_1^i|^2 + |\delta - \bar{\delta}|^2\right) \geqq 0,$$

for all $\bar{\delta} \geqq 0$ and points $\{\bar{\alpha}_0^i, \bar{\alpha}_1^i, \bar{\gamma}_0^i, \bar{\gamma}_1^i\}$ in M. We conclude from this inequality, along with [2, Props. 2.3.2, 2.4.2], that $\lambda \geqq 0$ and the inclusion (5.8) holds.

We now return to (5.9). Set $(\bar{\alpha}_0^j, \bar{\alpha}_1^j, \bar{\gamma}_0^j, \bar{\gamma}_1^j) = (\alpha_0^j, \alpha_1^j, \gamma_0^j, \gamma_1^j)$, $t_0^j = \tau_0^j$ and $t_1^j = \tau_1^j$ for all j, $1 \leqq j \leqq k$, and set $\bar{\delta} = \delta$. Select an integer i, $1 \leqq i \leqq k$, and set $y_j(\cdot) = z_j(\cdot)$ for all $j \neq i$. We deduce that $z_i(\cdot)$ solves the free-endpoint, fixed time problem of minimizing

$$\lambda g_i \cdot y(\tau_1^i) + s_0^i \cdot y(\tau_0^i) - s_1^i \cdot y(\tau_1^i)$$

$$+ m[|g_i \cdot y(\tau_1^i) - g_i \cdot z_i(\tau_1^i)|^2 + |y(\tau_0^i) - z_i(\tau_0^i)|^2 + |y(\tau_1^i) - z_i(\tau_1^i)|^2]$$

over component trajectories $y(\cdot): [\tau_0^i, \tau_1^i] \to \mathbb{R}^{n_i}$.

Suppose first that $\tau_0^i \neq \tau_1^i$. Since graph $\{z_i(\cdot)\}$ is assumed to be interior to $\bar{\Gamma}_i$ we can apply known necessary conditions [2, Thm. 3.2.6] to this problem, and conclude existence of an absolutely continuous function $p_i(\cdot)$ satisfying (5.3), (5.4), and (5.5). If $\tau_0^i = \tau_1^i$ (=: τ^i), the minimizing property of $z_i(\cdot)$ implies that $\lambda g_i + s_0^i - s_1^i = 0$. Otherwise expressed, there exists a vector that we write as $p_i(\tau^i)$, such that $p_i(\tau^i) = s_0^i$ and $p_i(\tau^i) = s_1^i - \lambda g_1$. Thus (5.4) and (5.5) are verified. In this case (5.3) is trivially satisfied.

There remains (5.6) and (5.7). Select an integer i, $1 \leqq i \leqq k$. It is convenient to separate the cases in which the time interval $[\tau_0^i, \tau_1^i]$ is degenerate and nondegenerate. Suppose first that $\tau_1^i > \tau_0^i$. Since z_i is assumed to have graph interior to $\tilde{\Gamma}^i$, we may choose $t_1^i \in I_i$ such that $t_1^i > \tau_1^i$. We proceed to extend $z_i(\cdot)|_{[\tau_0^i, \tau_1^i]}$ to $[\tau_0^i, t_1^i]$ thereby defining a component trajectory $y_i(\cdot)$. The hypotheses are satisfied under which Aumann's Selection Theorem (see [1]) applies to yield an absolutely continuous function $\xi: [\tau_1^i, t_1^i] \to \mathbb{R}^{n_i}$ such that $\xi(\tau_1^i) = z_i(\tau_1^i)$ and whose derivative has the following selection property:

$$\dot{\xi}(t) \in F_i(t, z_i(\tau_1^i)) \cap E_i(t) \quad \text{a.e.}$$

Here

$$E_i(t) = \{e \,|\, p_i(\tau_1^i) \cdot e = \max [p_i(\tau_1^i) \cdot e' \,|\, e' \in F_i(t, z_i(\tau_1^i))]\}.$$

Then

$$p_i(\tau_1^i) \cdot \dot{\xi}(t) = \mathcal{H}_i(t, z(\tau_1^i), p_i(\tau_1^i)) \quad \text{a.e. } t \in [\tau_1^i, t_1^i]. \tag{5.10}$$

It now follows from Theorem [2] of 3.1.6 and the hypotheses on $F_i(\cdot, \cdot)$ that there exists an absolutely continuous function $\xi(\cdot): [\tau_1^i, t_1^i] \to \mathbb{R}^{n_i}$ for which the following is true:

$$\begin{aligned} &\dot{\xi}(t) \in F_i(t, \xi(t)) \quad \text{a.e. } t \in [\tau_1^i, t_1^i], \\ &\xi(t_1^i) = z_i(\tau_1^i), \\ &[t_1^i - \tau_1^i]^{-1} \int_{\tau_1^i}^{t_1^i} |\dot{\xi}(s) - \dot{\bar{\xi}}(s)| \, ds \leqq K^2 \exp\{K(t_1^i - \tau_1^i)\}(t_1^i - \tau_1^i). \end{aligned} \tag{5.11}$$

Note that this bound has limit zero as $t_1^i \downarrow \tau_1^i$.

Now examine inequality (5.9) in the following situation. We take $\bar{\delta} = \delta$ and $\{\bar{\alpha}_0^j, \bar{\alpha}_1^j, \bar{\gamma}_0^j, \bar{\gamma}_1^j\} = \{\alpha_0^j, \alpha_1^j, \gamma_0^j, \gamma_1^j\}$ for all j, $1 < j \leqq k$. For $j \neq i$ take $(t_0^j, t_1^j, y_j(\cdot)) = (\tau_0^j, \tau_1^j, z_j(\cdot))$. Take also $t_0^i = \tau_0^i$ and define $y_i(\cdot) : [\tau_0^i, t_1^i] \to \mathbb{R}^{n_i}$ to be

$$y_i(t) = \begin{cases} z_i(t) & \text{for } t \in [\tau_0^i, \tau_1^i], \\ \xi(t) & \text{for } t \in [\tau_1^i, t_1^i]. \end{cases}$$

Write $\varepsilon' = t_1^i - \tau_1^i$ and divide across by ε'. There results

$$h_1^i - (s_1^i - \lambda g_i)\left((\varepsilon')^{-1} \int_{\tau_1^i}^{\tau_1^i + \varepsilon'} \dot{\xi}(s)\, ds\right) + (\varepsilon')^{-1} m\Delta \geqq 0.$$

Since $p_i(\tau_1^i) = s_1^i - \lambda g_i$, and by (5.10),

$$-h_1^i + (\varepsilon')^{-1} \int_{\tau_1^i}^{\tau_1^i + \varepsilon'} \mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i))\, dt$$

$$\leqq (\varepsilon')^{-1} m\Delta + (\varepsilon')^{-1} p_i(\tau_1^i) \cdot \int_{\tau_1^i}^{\tau_1^i + \varepsilon'} |\dot{\xi}(s) - \dot{\bar{\xi}}(s)|\, ds.$$

Because of the bound on F_i we have $|y_i(t_1^i) - z_i(\tau_1^i)| \leqq K|t_1^i - \tau_0^i|$, from which it follows that $\Delta/\varepsilon' \to 0$ as $\varepsilon' \downarrow 0$. The second term on the right-hand side is zero in the limit, by (5.11). It follows that

$$\limsup_{\varepsilon' \downarrow 0} (\varepsilon')^{-1} \int_{\tau_1^i}^{\tau_1^i + \varepsilon'} [\mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)) - h_1^i]\, dt \leqq 0.$$

This means that

$$h_1^i \in \operatorname*{ess}_{t \to \tau_1^i} \mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)) + [0, \infty), \tag{5.12}$$

since otherwise $\mathscr{H}_i - h_1^i > 0$ almost everywhere on some neighbourhood of τ_1^i, in contradiction of the inequality. Similar reasoning, but where we now choose $t_1^i < \tau_1^i$, produces

$$\liminf_{\varepsilon' \downarrow 0} (\varepsilon')^{-1} \int_{\tau_1^i - \varepsilon'}^{\tau_1^i} [\mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)) - h_1^i]\, dt \geqq 0$$

from which it follows that

$$h_1^i \in \operatorname*{ess}_{t \to \tau_1^i} \mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)) + (-\infty, 0]. \tag{5.13}$$

The two inclusions (5.12) and (5.13) imply

$$h_1^i \in \operatorname{co} \operatorname*{ess}_{t \to \tau_1^i} \mathscr{H}_i(t, z_i(\tau_1^i), p_i(\tau_1^i)).$$

The same arguments applied to the left endtime τ_0^i of the ith component trajectory $x_i(\cdot)$ yield

$$h_0^i \in \operatorname{co} \operatorname*{ess}_{t \to \tau_0^i} \mathscr{H}_i(t, z_i(\tau_0^i), p_i(\tau_0^i)).$$

We have proved (5.6) and (5.7) in the case $\tau_1^i > \tau_0^i$. To deal with the remaining case we suppose that $\tau_1^i = \tau_0^i$ $(=: \tau^i)$. Those aspects of earlier reasoning concerned with extending component trajectories to left or right remain valid in the degenerate case and yield the inequalities

$$\limsup_{\varepsilon' \downarrow 0} (\varepsilon')^{-1} \int_{\tau^i}^{\tau^i + \varepsilon'} [\mathscr{H}_i(t, z_i(\tau^i), p_i(\tau^i)) - h_1^i]\, dt \leqq 0,$$

$$\liminf_{\varepsilon' \downarrow 0} (\varepsilon')^{-1} \int_{\tau^i - \varepsilon}^{\tau^i} [\mathscr{H}_i(t, z_i(\tau^i), p_i(\tau^i)) - h_0^i]\, dt \geqq 0.$$

We will be able to conclude the desired inclusions (5.6) and (5.7) from these inequalities provided we can show that $h_1^i = h_0^i$. To this end we return to inequality (5.9). We set $\bar{\delta} = \delta$, and, for $j \neq i$, $\{\bar{\alpha}_0^j, \bar{\alpha}_1^j, \bar{\gamma}_0^j, \bar{\gamma}_1^j\} = \{\alpha_0^j, \alpha_1^j, \gamma_0^j, \gamma_1^j\}$ and $y_j(\cdot) = z_j(\cdot)$. We also set $t_1^i = \tau^i + \varepsilon'$ for some small nonzero ε' and $y_i(\cdot) \equiv z_i(\tau^i)$. Dividing across by ε' and passing to the limit as $\varepsilon' \to 0$, both from above and below, we obtain $0 \leqq -h_0^i + h_1^i \leqq 0$, i.e., $h_1^i = h_0^i$, as required. Proof of the lemma is complete.

The next step of the proof is to regard the proximal normal $[\{h_0^i, -h_1^i, -s_0^i, s_1^i\}, -\lambda]$ of Lemma 5.2, and its base point $[\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}, V(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}) + \delta]$ in epi V, as general terms in sequences such that

$$\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\} \to \{0, 0, 0, 0\} \quad \text{and} \quad V(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\}) + \delta \to V(\{0, 0, 0, 0\}).$$

Let $\{\tau_0^i, \tau_1^i, z_i(\cdot)\}$ be a solution to $Q(\{\rho_0^i, \rho_1^i, \sigma_0^i, \sigma_1^i\})$. We can arrange by subsequence extraction that $\tau_0^i \to T_0^i$, $\tau_1^i \to T_1^i$ and $z_i(\cdot) \to x_i(\cdot)$ uniformly, and also that the graph of $z_i(\cdot)$ is interior to $\tilde{\Gamma}^i$ along the sequence. Now apply Lemma 5.1. It follows from (5.4), (5.6), and (5.8) and hypotheses (I4) and (I5) that

$$\{-h_0^i, h_1^i, p_i(\tau_0^i), -p_i(\tau_1^i)\} \in \lambda \partial g + c\left[\sum_i |p_i(T_1^i)| + \lambda\right] \partial d_M. \tag{5.14}$$

Here ∂g and ∂d_Λ are evaluated at the point

$$\{\tau_0^i - \rho_0^i, \tau_1^i - \rho_1^i, z_i(\tau_0^i) - \sigma_0^i, z_i(\tau_1^i) - \sigma_1^i\}.$$

The magnitude of the number c is determined solely by the constant K of hypotheses (I4) and (I5) and by $\sum_i |g_i|$. We readily deduce from the fact that the proximal normal vector $[\{h_0^i, -h_1^i, -s_0^i, s_1^i\}, -\lambda]$ is nonzero that $[\sum_i |p_i(\tau_1^i)| + \lambda]$ is nonzero. Replace $\{p_i(\cdot)\}$ and λ by scaled versions, and so arrange that $[\sum_i |p_i(\tau_1^i)| + \lambda] = 1$. We thereby render the $p_i(\cdot)$'s elements in uniformly bounded and equicontinuous families of functions. By subsequence extraction we can then arrange that each $p_i(\cdot)$ has uniform limit a Lipschitz continuous function on I_i. We can also arrange that each $x_i(\cdot)$ has uniform limit $x_i(\cdot)$ (we appeal to Lemma 5.1 at this point) and the bounded sequences with general terms h_0^i, h_1^i and λ have limits also.

The differential inclusion (5.3) is preserved in the limit (see [2, Thm. 3.1.7]) along with the transversality condition (5.14) (by the upper semicontinuity properties of generalized gradients), in which component processes $z_i(\cdot)$ are replaced by $x_i(\cdot)$ and the generalized gradients ∂g and ∂d_M are evaluated at $\{T_0^i, T_1^i, x(T_0^i), x(T_1^i)\}$. Clearly we have $[\sum_i |p_i(T_1^i)| + \lambda] = 1$ in the limit. Finally we note that, since $z_i(\tau_1^i) \to x_i(T_1^i)$ and $p_i(\tau_1^i)$ converges to the value of the limiting costate function at T_1^i and in view of Lemma 2.1, the interpretation (5.6) and (5.7) of the h_0^i's and h_1^i's as convex essential values continue to hold good in the limit. This concludes proof of Theorem 4.1 in the presence of the supplementary hypotheses (IU) and (IL).

Removal of (IL). We next prove the special case of Theorem 4.1 in which the data is assumed to satisfy hypotheses (I1)–(I5) and (IU), but possibly not (IL).

We add an additional scalar-valued component trajectory $z(\cdot)$ and supplement the former dynamical equations as follows:

$$\dot{y}_i \in F_i(t, y_i) \quad \text{a.e. on } [\tau_0^i, \tau_1^i]$$

for $i = 1, \ldots, k$, and

$$\dot{z} \in \{0\}, \quad \text{a.e. on } [\sigma_0, \sigma_1].$$

The endpoints of the trajectories are constrained to satisfy

$$(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i), y_i(\tau_1^i)\}, (\sigma_0, \sigma_1, z(\sigma_0), z(\sigma_1)) \in \tilde{M}.$$

The set $\tilde{M}$ is taken to be

$$\tilde{M} := \{(a, 0, 1, z_0, z_1) | a \in M, z_0 \geqq g_e(a, 0, 1, z_1)\}$$

in which g_e is the extension of $\hat{g}$ to $\prod_{i=1}^{k} (\mathbb{R}\times\mathbb{R}\times\mathbb{R}^{n_i}\times\mathbb{R}^{n_i})\times(\mathbb{R}\times\mathbb{R}\times\mathbb{R})$ defined by

$$g_e(a, (\sigma_0, \sigma_1, z_1)) = g(a).$$

(Note in particular that the endtimes of the new component trajectory are fixed at $t=0$ and 1 and its value at the right endtime is unconstrained.) The new objective function is

$$\tilde{g}(a, (\sigma_0, \sigma_1, z_0, z_1)) = z_1.$$

It is a simple matter to see that this modified problem continues to satisfy the hypotheses (I1)-(I5) as well as (IU) and that it has a solution

$$(\{T_0^i, T_1^i, x_i(\cdot)\}, (0, 1, y(\cdot)) \equiv g(\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\})).$$

Since, in addition, the problem clearly satisfies (IL), the conclusions of the theorem for this problem are available to us. These are seen to imply existence of a number $\lambda \geqq 0$, numbers α, β, and q, and functions $p_i(\cdot)$ and numbers h_0^i, h_1^i for $i=1, \ldots, k$, such that $[\sum_i |p_i(T_1^i)| + q + \lambda] = 1$, the Hamiltonian inclusions are satisfied, the h_0^i, h_1^i's are convex essential values of the Hamiltonian functions, and

$$(5.15) \quad \{-h_0^i, h_1^i, p_i(T_0^i), -p_i(T_1^i)\}, (\alpha, \beta, -q, q) \in c\,\partial d_{\tilde{M}} + \lambda[0, (0, \cdots, 0, 1)].$$

Here c is a number whose magnitude is governed by the constant K of hypotheses (I4) and (I5). $\partial d_{\tilde{M}}$ is evaluated at $(\bar{a}, (0, 1, g(\bar{a}), g(\bar{a})))$ in which $\bar{a}$ denotes $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$.

We pause to take note of the following estimate on points in the generalized gradient of a distance function (see [4, Lemma 4.1]).

LEMMA 5.3. *Let $S \subset \mathbb{R}^k$ be a closed set and take a point $\bar{s} \in S$. Suppose there is a constant $\delta > 0$ and a function $l: \bar{s} + \delta B \to \mathbb{R}$ such that l is Lipschitz continuous of rank at most K_l on $\bar{s} + \delta B$. Then for all $R \geqq (1 + K_l^2)^{1/2}$ we have*

$$\partial d_{\mathrm{epi}(l+\psi_S)}(\bar{s}, l(\bar{s})) \subset \{(\zeta, -\varepsilon) | \zeta \in \varepsilon\partial l(\bar{s}) + R\,\partial d_S(\bar{s}), \varepsilon \geqq 0\}.$$

Here $\psi_S(s)$ equals zero if $s \in S$, $+\infty$ otherwise.

Appealing to this lemma, and also to the fact that

$$(5.16) \qquad \partial d_{M\times\{0\}\times\{1\}\times\mathbb{R}} \subset \partial d_M \times B \times B \times \{0\}$$

(see [2, Thm. 2.5.6]), we deduce from (5.15) that $q = \lambda$ and

$$\{-h_0^i, h_1^i, p_i(T_0^i), -p_i(T_1^i)\} \in \lambda\,\partial g + c(1 + \bar{K}^2)^{1/2}\,\partial d_M.$$

Here $\bar{K}$ is an upper bound on the Lipschitz rank of g on a neighbourhood of $\{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$. It remains to replace the $p_i(\cdot)$'s and λ by scaled versions which have the property that $\sum_i |p_i(T_1^i)| + \lambda = 1$, as is possible since $\lambda = q$. The Hamiltonian inclusions are unaffected. As for the transversality condition, this is clearly valid provided $c(1 + \bar{K}^2)^{1/2}$ is replaced by $2c(1 + \bar{K}^2)^{1/2}$, a number whose magnitude is governed by the constant K of hypotheses (I4) and (I5) and by $\bar{K}$.

Removal of (IU). Suppose finally that the data of the problem satisfy hypotheses (I1)-(I5), but possibly not (IU). As usual $\{T_0^i, T_1^i, x_i(\cdot)\}$ is the solution to (P) under consideration.

Consider a modified version of the problem in which each component trajectory has its state dimension increased by one and is now governed by the dynamics

$$\dot{z}_i = (y_i - x_i(t))^2 \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i],$$
$$\dot{y}_i \in F_i(t, y_i) \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i].$$

The objective functional is now taken to be $\tilde{g}$:

$$\tilde{g}(\{\tau_0^i, \tau_1^i, (z_0^i, y_0^i), (z_1^i, y_1^i)\}) = g(\{\tau_0^i, \tau_1^i, y_0^i, y_1^i\}) + \sum_i (|z_i(\tau_1^i)|^2 + |\tau_0^i - T_0^i|^2 + |\tau_1^i - T_1^i|^2)$$

and the original constraint set M is replaced by $\tilde{M}$:

$$\tilde{M} = \{\{\tau_0^i, \tau_1^i, (z_0^i, y_0^i), (z_1^i, y_1^i)\} | \{\tau_0^i, \tau_1^i, y_0^i, y_1^i\} \in M \text{ and } z_0^i = 0 \text{ for } i = 1, \cdots, k\}.$$

Note that $\{T_0^i, T_1^i, (z_i(\cdot) \equiv 0, x_i(\cdot))\}$ is a solution to this problem, and furthermore this solution is unique because we have arranged that deviations from it are penalized. It is clear that the modified problem satisfies (IU) in addition to hypotheses (I1)-(I5). We are permitted then to apply the special case of the theorem already proved. The conclusions of the theorem for the original problem are seen to follow. (Presence of squared terms in the modified problem ensures that the additional dynamics and perturbations to the objective functional do not impinge on the necessary conditions.)

6. Proof of Theorem 3.1. Our proof of Theorem 3.1, by application of Theorem 4.1 to an auxiliary problem, has much in common with the proof of the maximum principle of Clarke for boundary points of the attainable set [2, pp. 201-209]. We enter into the details of the argument only when there is a significant departure, most notably in the setting up of the auxiliary problem.

Choose $\varepsilon > 0$ such that

$$\text{graph } \{x_i(\cdot)\} + 2\varepsilon B \subset X^i, \qquad i = 1, \cdots, k,$$

and define the sets $\tilde{X}^i$, $i = 1, \cdots, k$, to be

$$\tilde{X}^i = \text{graph } \{x_i(\cdot)\} + \varepsilon \bar{B}, \qquad i = 1, \cdots, k.$$

Now let I_i denote the interval $[T_0^i - \varepsilon, T_1^i + \varepsilon]$, $i = 1, \cdots, k$.

A point of terminology: An "extended multiprocess" is just a multiprocess $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ with $[\tau_0^i, \tau_1^i] \subset I_i$ and graph $\{y_i(\cdot)\} \subset \tilde{X}_i$, $i = 1, \cdots, k$, in all respects except $y_i(\cdot)$ is viewed as a function with domain I_i, obtained from the original multiprocess by constant extrapolation to left and right, and $w_i(\cdot)$ is now taken to represent an equivalence class of functions equal almost everywhere, $i = 1, \cdots, k$. Denote by W the set of extended processes that satisfy $\{\tau_0^i, \tau_1^i, y_i(\tau_0^i)\} \in C$. Let $\Delta: W \times W \to \mathbb{R}$ be the function

$$\Delta(\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}, \{\bar{\tau}_0^i, \bar{\tau}_1^i, \bar{y}_i(\cdot), \bar{w}_i(\cdot)\})$$
$$:= \sum_i [|\tau_0^i - \bar{\tau}_0^i| + [|\tau_1^i - \bar{\tau}_1^i| + |y_i(\tau_0^i) - \bar{y}_i(\bar{\tau}_0^i)|$$
$$+ \mathscr{L}\text{-meas } \{t \in [\tau_0^i \vee \bar{\tau}_0^i, \tau_1^i \wedge \bar{\tau}_1^i] | w_i(t) \neq \bar{w}_i(t)\}].$$

Here $a \vee b$, $a \wedge b$ denote the maximum and minimum of a and b, respectively. Simple modifications of the proof of Lemma 1 of [2, p. 202] establish the following lemma.

LEMMA 6.1. *The function Δ is a metric on W, and (W, Δ) is a complete metric space. Let $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ represent the general term in a sequence of points in (W, Δ) converging to a point $\{\bar{\tau}_0^i, \bar{\tau}_1^i, \bar{y}_i(\cdot), \bar{w}_i(\cdot)\}$. Then* $\limsup_{t \in I_i} |y_i(t) - \bar{y}_i(t)| = 0$, *for* $i = 1, \cdots, k$.

Let n be a positive integer and let ζ be a point in $\psi(\{x_i(T_1^i)\})+n^{-2}B$ such that $\zeta \notin \mathscr{R}_{\psi,C}$, and define the function $F:(W,\Delta)\to\mathbb{R}$ to be

$$F(\{\tau_0^i, \tau_1^i, y_i(\cdot), u_i(\cdot)\}) := |\zeta - \psi(\{y_i(\tau_1^i)\})|.$$

By Lemma 6.1, F is a continuous function. We have

$$F(\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}) < \inf_{e\in W} F(e) + n^{-2}.$$

(There is a minor abuse of notation here that we shall repeat. The same symbols are used for the solution $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ and the corresponding *extended* multiprocess.) The hypotheses are met under which Ekeland's Theorem [6] is applicable. This tells us that there exists a point $\bar{e}=\{\bar{T}_0^i, \bar{T}_1^i, \bar{x}_i(\cdot), \bar{u}_i(\cdot)\}$ in W such that, writing $e=\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$, we have

$$\Delta(e,\bar{e}) \leqq n^{-1} \tag{6.1}$$

$$F(\bar{e}) \leqq F(e') + n^{-1}\Delta(e',\bar{e}) \quad \text{for all } e'\in W. \tag{6.2}$$

In view of (6.1), we may arrange that

$$\operatorname{graph}\{\bar{x}_i(\cdot)\} + (\varepsilon/2)B \subset \tilde{X}^i, \qquad i=1,\cdots,k$$

by choosing n sufficiently large. The following lemma is then a consequence of the minimizing property (6.2) of $\bar{e}$.

LEMMA 6.2. *Let* $\{\tau_0^i, \tau_1^i, y_i(\cdot), w_i(\cdot)\}$ *be any multiprocess such that* $\{\tau_0^i, \tau_1^i, y_i(\tau_0^i)\}\in C$ *and*

$$\sup_{t\in I_i} |y_i(t)-\bar{x}_i(t)| \leqq \frac{\varepsilon}{2}$$

for $i=1,\cdots,k$. *Then*

$$|\zeta-\psi(\{y_i(\tau_1^i)\})| + n^{-1}\sum_i \left([(\tau_0^i-\bar{T}_0^i)\vee 0] + [(\bar{T}_1^i-\tau_1^i)\vee 0] + |y(\tau_0^i)-\bar{x}(\tau_0^i)| + \int_{\tau_0^i}^{\tau_1^i} \mathscr{X}_i(t, w_i(t))\,dt \right) \geqq |\zeta-\psi(\{\bar{x}(T_1^i)\})|.$$

Here

$$\mathscr{X}_i(t,w) := \begin{cases} 1 & \text{if } t\notin [T_0^i, T_1^i] \text{ or } w\neq \bar{u}_i(t), \\ 0 & \text{otherwise.} \end{cases}$$

We proceed to interpret Lemma 6.2 in such a way that Theorem 4.1 becomes applicable. This requires us to consider new component state vectors $Y_i=(z_i, y_i)$ and associated differential inclusions with right-hand sides

$$F_i(t, Y_i) := \{[\mathscr{X}_i(t,w), \phi(t,y,w)] \mid w\in U_i^i\}.$$

We take the endpoint constraint set Λ to be

$$\Lambda := \{\{\tau_0^i, \tau_1^i, (z_0^i, y_0^i), (z_1^i, y_1^i)\} \mid \{\tau_0^i, \tau_1^i, y_0^i\}\in C,\ z_0^i=0,\ i=1,\cdots,k\}$$

and the function g to be

$$g(\{\tau_0^i, \tau_1^i, Y_0^i, Y_1^i\}) := |\zeta-\psi(\{y_1^i\})| + n^{-1}\sum_i (z_1^i + (\tau_0^i-\bar{T}_0^i)\vee 0 + (\bar{T}_1^i-\tau_1^i)\vee 0 + |y_0^i-\bar{x}(\bar{T}_0^i)|)$$

in which $Y_0^i=(z_0^i, y_0^i)$ and $Y_1^i=(z_1^i, y_1^i)$.

It is a simple matter to deduce from Lemma 6.2 that $(\bar{T}_0^i, \bar{T}_1^i, (\bar{z}(t) \equiv \int_0^t \mathscr{X}(s, \bar{u}(s))\, ds, \bar{x}(\cdot)))$ is a solution to the following optimization problem, which we label (P(n)):

(P(n))
$$\text{minimize} \quad g(\{\tau_0^i, \tau_1^i, Y_i(\tau_0^i), Y_i(\tau_1^i)\})$$
$$\text{subject to}$$

(6.3)
$$\dot{Y}_i(t) \in F_i(t, Y_i(t)) \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i],$$
$$\text{graph}\,\{Y_i(\cdot)\} \subset \mathbb{R} \times (\text{graph}\,\{\bar{x}(\cdot)\} + (\varepsilon/2)B), \qquad i = 1, \cdots, k,$$

(6.4)
$$\{\tau_0^i, \tau_1^i, Y_i(\tau_0^i), Y_i(\tau_1^i)\} \in \Lambda.$$

From this point the proof follows closely from [2, pp. 205–209] and we merely outline what is involved. We impose for the time being the following supplementary hypothesis.

(HF)
$$U_t^i \text{ is a finite set, for } t \in \mathbb{R},\ i = 1, \cdots, k.$$

Observe that condition (6.4) leaves $\{Y_i(\tau_1^i)\}$ unconstrained. We deduce by means of standard arguments (see, e.g., [2, p. 117]) that elements $\{Y_i(\cdot)]$ in problem (P(n)) that satisfy (6.3) and (6.4) can be approximated uniformly by elements associated with the differential inclusions

(6.5)
$$\dot{Y}_i(t) \in \text{co}\, F_i(t, Y_i(t)) \quad \text{a.e. } t \in [\tau_0^i, \tau_1^i],$$

which satisfy (6.4). (Hypothesis (HF) is significant at this point, since it ensures that F_i has closed values.) Because g is a continuous function, we can conclude that $\{\bar{T}_0^i, \bar{T}_1^i, \bar{z}_i(\cdot), \bar{x}_i(\cdot)\}$ is also a solution to a new problem, denoted by co (P(n)), in which condition (6.3) in (P(n)) is replaced by its convexification (6.5).

Problem co (P(n)) meets the requirements for application of Theorem 4.1. Arguing as in [2], we show that for each n sufficiently large there exist functions $\{p_i(\cdot): I_i \to \mathbb{R}^{n_i}\}$ (which are uniformly bounded and equicontinuous with respect to the index n), a vector v of unit length and measurable sets $A_i(n) \subset [T_0^i, T_1^i]$ such that

(6.6)
$$\mathscr{L} - \text{meas}\,\{A_i(n)\} \to T_1^i - T_0^i \quad \text{as } n \to \infty,$$
$$p_i(\bar{T}_1^i) \in \partial\psi^*(\{\bar{x}_i(\bar{T}_1^i)\})v$$

and for all $t \in A_i(n)$

(6.7)
$$-\dot{p}_i(t) \in \partial_x H(t, \bar{x}_i(t), u_i(t), p_i(t))$$

(6.8)
$$H_i(t, \bar{x}_i(t), u_i(t), p_i(t)) \geqq \max_{w \in U_t^i} \{H_i(t, \bar{x}_i(t), w, p_i(t))\} - n^{-1}.$$

(The hypothesis (HF) is once again involved.)

We can also show that

(6.9)
$$\{-h_0^i, h_1^i, p_i(\bar{T}_0^i)\} \in \bar{K}\partial_C(\{\bar{T}_0^i, \bar{T}_1^i, \bar{x}_i(\bar{T}_0^i)\}) + n^{-1}\bar{K}B$$

for some elements $\{h_0^i, h_1^i\}$ that satisfy

(6.10)
$$h_0^i \in \text{co} \operatorname*{ess}_{t \to T_1^i} \mathscr{h}_i(t, \bar{x}_i(\bar{T}_0^i), p_i(\bar{T}_1^i)) + n^{-1}\bar{K}B,$$

(6.11)
$$h_1^i \in \text{co} \operatorname*{ess}_{t \to T_1^i} \mathscr{h}_i(t, \bar{x}_i(\bar{T}_1^i), p_i(\bar{T}_1^i)) + n^{-1}\bar{K}B.$$

Here $\bar{K}$ is a number whose magnitude is governed by the constant K of hypotheses (H3) and (H4) and by the Lipschitz rank of ψ in a neighbourhood of $\{x_i(T_1^i)\}$. The function $\mathscr{h}_i$ is

$$\mathscr{h}_i(t, x, p) := \max_{w \in U_t^i} H_i(t, x, w, p).$$

We now extract subsequences to ensure that the h_0^i's, h_1^i's and v have limits and the $p_i(\cdot)$'s have uniform limits as $n \to \infty$ (we retain the symbols h_0^i, h_i^i, v, $p_i(\cdot)$ for the limits), and that

$$A := \bigcup_n \bigcap_{n \leqq m} A_i(m)$$

has measure $T_1^i - T_0^i$. The $\bar{x}_i(\cdot)$'s converge uniformly to $x_i(\cdot)$ by (6.1), as $n \to \infty$. Appealing to Theorem 3.1.7 of [2] and the upper semicontinuity properties of generalized gradients, we can take the limit, $n \to \infty$, in (6.7) and (6.8) for all $t \in A$, and in (6.6) and (6.7). Passing to the limit in (6.10) and (6.11) is justified by Lemma 2.1. There result the assertions of Theorem 3.1.

We recall that the theorem has been proved under the supplementary hypothesis (HF); this is disposed of by a componentwise application of techniques on p. 207 of [2].

Proof of Theorem 3.2. Let $\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}$ solve (P). Consider multiprocesses with state vectors $((y_1, \cdots, y_k), (\bar{y}_1, \cdots, \bar{y}_k), z) \in (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}) \times (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}) \times \mathbb{R}$. The dynamical equations and associated constraints are now taken to be

$$\begin{aligned}
&\dot{y}_i = f_i(t, y_i, u_i) && \text{a.e. } [\tau_0^i, \tau_1^i],\\
&u_i(t) \in U_t^i && \text{a.e. } [\tau_0^i, \tau_1^i],\\
&y_i(t) \in X_t^i && \text{for all } [\tau_0^i, \tau_1^i],\\
&\dot{\bar{y}}_i = \bar{u}_i && \text{a.e. } [\sigma_0^i, \sigma_1^i],\\
&\bar{u}_i(t) \in \{0\} && \text{a.e. } [\sigma_0^i, \sigma_1^i],
\end{aligned}$$

for $i = 1, \cdots, k$, and

$$\begin{aligned}
&\dot{z} = w && \text{a.e. } [\sigma_0^i, \sigma_1^i],\\
&w(t) \in \{0\} && \text{a.e. } [\sigma_0^i, \sigma_1^i].
\end{aligned}$$

The endpoints constraint of interest is

$$(\{\tau_0^i, \tau_1^i, y_i(\tau_0^i)\}, \{\sigma_0^i, \sigma_1^i, \bar{y}_i(\sigma_0^i)\}, \{\sigma_0, \sigma_1, z(\sigma_0)\}) \in C^+$$

where C^+ is the set

$$C^+ := S_1 \cap S_2$$

in which

$$\begin{aligned}
S_1 &:= \{(\{\tau_0^i, \tau_1^i, x_0^i\}, \{0, 1, \bar{x}_0^i\}, (0, 1, z_0)) | \{\tau_0^i, \tau_1^i, x_0^i, \bar{x}_0^i\} \in \Lambda\},\\
S_2 &:= \{(\{\tau_0^i, \tau_1^i, x_0^i\}, \{\sigma_0^i, \sigma_1^i, \bar{x}_0^i\}, (\sigma_0, \sigma_1, z_0)) | (\{\tau_0^i, \tau_1^i, x_0^i, \bar{x}_0^i\}, z_0) \in \operatorname{epi} f\}.
\end{aligned}$$

Now consider the mapping

$$\psi^+ : (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}) \times (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}) \times \mathbb{R} \to (\mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}) \times \mathbb{R}$$

defined by

$$\psi^+(\{x_i\}, \{\bar{x}_i\}, z) = (\{x_i - \bar{x}_i\}, z).$$

It can be shown by means of a simple contradiction argument that

$$\begin{aligned}
&(\{T_0^i, T_1^i, x_i(\cdot), u_i(\cdot)\}, \{0, 1, \bar{x}_i(\cdot) \equiv x_i(T_1^i), \bar{u}_i(\cdot) \equiv 0\},\\
&(0, 1, z(\cdot) \equiv f(\{T_0^i, T_1^i, x_i(T_0^i, x_i(T_1^i)\}), w(\cdot) \equiv 0))
\end{aligned}$$

is a boundary multiprocess with respect to C^+ and ψ^+.

All hypotheses are satisfied under which Theorem 3.1 is applicable. Accordingly there exist real numbers q, d_0, d_1 and h_0^i, h_1^i, $\bar{h}_0^i$, $\bar{h}_1^i$ for $i=1,\cdots,k$, vectors $\bar{p}_i \in \mathbb{R}^{n_i}$, $i=1,\cdots,k$, absolutely continuous functions $p_i(\cdot):[T_0^i, T_1^i]\to\mathbb{R}^{n_i}$, $i=1,\cdots,k$, vectors $\{s_i\}$, a number λ, and a positive number $\bar{K}$ whose magnitude is governed solely by the constant K of hypothesis (H3) and (H4) such that

$$-\dot{p}_i(t)\in\partial_x H_i, \qquad H_i(t, x_i(t), u_i(t), p_i(t)) = \max_{w\in U_i^i} H_i(t, x_i(t), w, p_i(t)).$$

$-h_0^i$ and h_1^i are convex essential values of the Hamiltonian functions, for almost everywhere $t\in[T_0^i, T_1^i]$, $i=1,\cdots,k$,

$$\{-h_0^i, h_1^i, p_i(T_0^i)\}, \{-\bar{h}_0^i, \bar{h}_1^i, \bar{p}_i\}, (-d_0, d_1, q)\in \bar{K}\,\partial d_{C^+}(\{T_0^i, T_1^i, x_i(T_0^i)\}, \{0, 1, x_i(T_1^i)\}, (0, 1, f)) \tag{7.1}$$

where f is evaluated at $e := \{T_0^i, T_1^i, x_i(T_0^i), x_i(T_1^i)\}$,

$$(\{-p_i(T_1^i), -\bar{p}_i\}, -q) = (\{s_i, -s_i\}, \lambda), \tag{7.2}$$

and

$$|(\{s_i, -s_i\}, \lambda)| = 1. \tag{7.3}$$

Keeping in mind Lemma 5.3 and inclusion (5.16), we deduce from (7.1) that $q\leqq 0$ and

$$\{-h_0^i, h_1^i, p_i(T_0^i), \bar{p}_i\}\in -q\partial f + \bar{K}(1+K_f^2)^{1/2}\partial d_\Lambda.$$

Here K_f is the Lipschitz rank of f in a neighbourhood of e. We conclude from this inclusion, from (7.2), and from (7.3) that $\lambda\geqq 0$, and the p_i's and λ can be scaled so that

$$\{-h_0^i, h_1^i, p_i(T_0^i), -p_i(T_1^i)\}\in \lambda\partial f + c\partial d_\Lambda \qquad \sum_i |p_i(T_1^i)| + \lambda = 1.$$

Here $c = 2^{3/2}\bar{K}(1+\bar{K}^2)^{1/2}$, a number whose magnitude is governed by the constant K of hypotheses (H3) and (H4) and by the Lipschitz rank of f in a neighbourhood of e. Surveying these relationships, we see that the elements $\{p_i(\cdot)\}$ and λ have all the properties of the component costate functions and the cost multiplier listed in Theorem 3.2. This concludes the proof.

REFERENCES

[1] R. J. AUMANN, *Measurable utility and the measurable choice theorem*, La Décision, Actes Coll. Int. du CNRS, Aix-en-Provence, 1967, pp. 15-26.

[2] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[3] F. H. CLARKE AND P. D. LOEWEN, *The value function in optimal control: sensitivity, controllability and time optimality*, SIAM J. Control Optim., 24 (1986), pp. 243-263.

[4] F. H. CLARKE, P. D. LOEWEN, AND R. B. VINTER, *Differential inclusions with free time*, Ann. Inst. H. Poincaré, 5 (1988), pp. 573-593.

[5] F. H. CLARKE AND R. B. VINTER, *Applications of optimal multiprocesses*, SIAM J. Control Optim., this issue, pp. 1048-1071.

[6] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324-353.

[7] T. G. GAUVIN, *The generalized gradient of a marginal function in mathematical programming*, Math. Oper. Res., 4 (1979), pp. 458-463.

[8] D. S. HAGUE, *Solution of multiple arc problems by the steepest descent method*, in Recent Advances in Optimization Techniques, A. Lavi and T. P. Vogl, eds., John Wiley, New York, 1965, pp. 489-518.

[9] A. D. IOFFE, *Necessary conditions in nonsmooth optimization*, Math. Oper. Res., (1984), pp. 159-189.

[10] L. W. NEUSTADT, *Optimization: A Theory of Necessary Conditions*, Princeton University Press, Princeton, NJ, 1976.

[11] R. W. RISHEL, *An extended Pontryagin principle for control systems whose control laws contain measures*, SIAM J. Control Optim., 3 (1965), pp. 191-205.

[12] R. T. ROCKAFELLAR, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Stud., 17 (1982), pp. 28-66.
[13] R. B. VINTER AND F. M. F. L. PEREIRA, *A maximum principle for optimal processes with discontinuous trajectories*, SIAM J. Control Optim., 26 (1988), pp. 205-229.
[14] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[15] ———, *Variational problems with unbounded controls*, SIAM J. Control Optim., 3 (1966), pp. 424-438.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1092-1107, September 1989

011

ON THE MINIMUM PRINCIPLE FOR CONTROLLED DIFFUSIONS ON MANIFOLDS*

M. H. A. DAVIS† AND M. P. SPATHOPOULOS‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. This paper concerns optimal control of a family of nondegenerate diffusions evolving on a compact manifold. Feedback controls based on complete observations are used. The controlled process is constructed by the horizontal lifting technique of Eells and Elworthy, and a minimum principle is derived from a dynamic programming argument. It is shown that the adjoint variable appearing in the minimum principle, which here is a one-form valued stochastic process, can be evaluated by solving a "heat equation" on the manifold.

Key words. stochastic control, dynamic programming, minimum principle, stochastic differential equation, diffusion on manifolds

AMS(MOS) subject classifications. 93E20, 60H10

1. Introduction. Stochastic versions of the Pontryagin minimum principle of optimal control theory have been a topic of interest since the pioneering work of Kushner [13] in the mid-1960s. In the case of controlled nondegenerate diffusions, the subject is best approached via dynamic programming; the Bellman equation is a quasilinear second-order parabolic partial differential equation that has much smoother solutions than the first-order equation arising in deterministic control theory.[1] A complete treatment is given by Fleming and Rishel [8]. The adjoint variable in the stochastic minimum principle is seen to be $p_t := \partial V(t, x_t)/\partial x$, where $V(t, x)$ is the value function and x_t the state of the controlled process. p_t can be characterized further by using a result from stochastic flow theory, namely, that the solution $x_{t,s}(x)$ of a stochastic differential equation (SDE) with initial point $x_{s,s}(x) = x$ depends (almost surely) smoothly on x. If the cost function is the terminal cost $\mathbb{E}[\theta(x_{1,s}(x))]$ and $\mathbb{E}^*$ denotes expectation with respect to the measure of the optimally controlled process, then $V(s, x) = \mathbb{E}^*[\theta(x_{1,s}(x))]$ and hence

$$\frac{\partial}{\partial x} V(s, x) = \mathbb{E}^*\left[\sum_{i,j} \frac{\partial \theta}{\partial x^i} \frac{\partial}{\partial x^j} x^i_{1,s}(x)\right].$$

Since $\partial x^i_{1,s}(x)/\partial x^j$ is computed from the "linearized equations" corresponding to the controlled SDE, this shows that the stochastic adjoint variable p_t is just the conditional expectation, given the information available at time t, of its deterministic counterpart. There is a slight modification in the formulation of the linearized equations due to the fact that feedback rather than "open loop" controls are being used. All of this is explained by Haussmann [9], [10].

In [3] a pathwise solution for the nonlinear filtering problem of a diffusion process x_t specified by its generator on a manifold has been derived using the horizontal lifting technique of Eells and Elworthy [6]. Motivated by this technique we formulate the completely observable stochastic control problem in the setting of controlled diffusions on manifolds specified by their generators. There are two reasons for doing so. First,

* Received by the editors January 5, 1987; accepted for publication (in revised form) January 13, 1989.
† Department of Electrical Engineering, Imperial College, London SW7 2BT, United Kingdom.
‡ Division of Dynamics and Control, Strathclyde University, Glasgow G1 1XS, Scotland.
[1] Vinter [17] gives an up-to-date discussion of the deterministic case.

there are important problems, such as controlling the orientation of a rigid body, that are naturally formulated in this way, and indeed it seems surprising that use of geometric methods in control theory has so far largely been restricted to qualitative questions such as controllability. Second, we obtain a new interpretation of the adjoint variable $\partial V/\partial x$: it satisfies a form of the so-called "heat equation for tensor fields" involving an intrinsic operator on tensor fields known as the *de Rham-Kodaira Laplacian.*

There is some related work by Duncan [5]. Duncan defines the solutions of stochastic systems in Riemannian manifolds in virtually the same way as we do here—and independently of Eells and Elworthy [6]—and treats stochastic control by martingale methods in the manner of Davis and Varaiya [4]. This however leaves the adjoint variable only as an implicitly-defined object.

The paper is organized as follows. Section 2 is a preliminary section giving some geometric notions and outlining the construction of Brownian motion on a manifold by "horizontal lifting." The control problem is formulated in § 3 and the dynamic programming results that in $\mathbb{R}^d$ are the same as those of Fleming and Rishel [8] are given in § 4. The last two sections, §§ 5 and 6, are devoted to obtaining the characterization of the adjoint variable mentioned above. This is the main result of the paper, and is stated as Theorem 6.6. Appendix A gives the proof of a technical result concerning the Bellman equation on manifolds.

2. Preliminaries. All of the following information can be found in Boothby [1] and Ikeda and Watanabe [11, Chap. 5]. Throughout the paper the summation convention over repeated indices is used. The Stratonovich and Itô stochastic integrals are denoted $\beta \circ dw$ and $\beta\, dw$, respectively.

Let M be a compact[2] C^∞ manifold of dimension d. Denote by $T_x(M)$ and $T_x^*(M)$, respectively, the tangent and cotangent spaces at $x \in M$. The *tangent bundle* is $TM = \{(x, v)\colon x \in M,\ v \in T_x(M)\}$ and the *bundle of linear frames* is

$$\mathrm{GL}(M) = \{r = (x, e)\colon x \in M,\ e = [e_1, \cdots, e_d],$$
$$e_i \in T_x(M) \text{ and } [e_1, \cdots, e_d] \text{ is a basis for } T_x(M)\}.$$

A *Riemannian metric* is a C^∞ field $\Phi = \{\Phi_x, x \in M\}$ of positive-definite symmetric bilinear forms on TM. Then Φ_x defines an inner product on $T_x(M)$. The *orthonormal frame bundle* $\mathrm{O}(M)$ is defined as

$$\mathrm{O}(M) = \{r = (x, e) \in \mathrm{GL}(M)\colon \Phi_x(e_i, e_j) = \delta_{ij},\ i, j = 1 \cdots d\}.$$

Both $\mathrm{GL}(M)$ and $\mathrm{O}(M)$ are themselves C^∞ manifolds. Let $\mathfrak{X}(M)$ denote the set of vector fields on M, i.e., the set of functions $X : x \mapsto X(x) \in T_x(M)$ such that $Xf(x)$ is a C^∞ function for each $f \in C^\infty(M)$. A *Riemannian* or *Levi-Civita connection* is a mapping $\nabla : \mathfrak{X}(M) \times \mathfrak{X}(M) \to \mathfrak{X}(M)$, written $(X, Y) \mapsto \nabla_X Y$ such that (i) $\nabla_X Y$ is bilinear in X, Y; (ii) $\nabla_{fX+gY} = f\nabla_X + g\nabla_Y$; (iii) $\nabla_X(fY) = f\nabla_X Y + (Xf)\,Y$ (here $f, g \in C^\infty(M)$); (iv) $\nabla_X Y - \nabla_Y X = XY - YX$; and (v) $X\Phi(Y, Z) = \Phi(\nabla_X Y, Z) + \Phi(Y, \nabla_X Z)$. There is a uniquely determined Riemannian connection corresponding to each Riemannian metric. In local coordinates, with $X(x) = \alpha^i \partial/\partial x^i$, $Y(x) = \beta^i \partial/\partial x^i$ we have $\Phi_x(X, Y) = a_{ij}(x)\alpha^i\beta^j$ for some positive definite symmetric matrix $[a_{ij}(x)]$ and

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}^k \frac{\partial}{\partial x^k} \qquad \left(\partial_i = \frac{\partial}{\partial x^i}\right)$$

[2] Undoubtedly, all results in this paper will extend to noncompact manifolds if some uniform non-explosion condition is imposed.

where the coefficients Γ_{ij}^k (the *Christoffel symbols*) are given by

$$\Gamma_{ij}^k = \frac{1}{2}\left(\frac{\partial}{\partial x^i} a_{mj} + \frac{\partial}{\partial x^j} a_{im} - \frac{\partial}{\partial x^m} a_{ij}\right) a^{km}$$

($[a^{km}] = [a_{km}]^{-1}$). An element $v \in T_r(\mathrm{O}(M))$ takes the form

$$v = \alpha^i \frac{\partial}{\partial x^i} + \beta_j^k \frac{\partial}{\partial e_j^k}$$

where $r = (x, e) \in \mathrm{O}(M)$ and $e_i = e_i^j \partial/\partial x^j$.

The connection defines a d-dimensional *horizontal subspace* H_r of $T_r(\mathrm{O}(M))$ as follows:

$$H_r = \left\{ v = \alpha^i \frac{\partial}{\partial x^i} - \Gamma_{kl}^i(x) e_j^l \alpha^k \frac{\partial}{\partial e_j^i} : (\alpha^i) \in R^d \right\}.$$

We denote by $\pi : \mathrm{O}(M) \to M$ the projection map $\pi r = x$. If $Y \in T_x(M)$, then $\tilde{Y} \in T_r(\mathrm{O}(M))$ is the *horizontal lift* of Y if $x = \pi r$, $\tilde{Y} \in H_r$, and $T_\pi \tilde{Y} = Y$, where $T_\pi \tilde{Y} f(r) = \tilde{Y}(f \circ \pi)(r)$. Any $X \in \mathfrak{X}(M)$, $X = X^i(x)(\partial/\partial x^i)$, has a unique horizontal lift to a vector field $\tilde{X} \in \mathfrak{X}(\mathrm{O}(M))$ defined by

$$\tilde{X}(x, e) = X^i(x) \frac{\partial}{\partial x^i} - \Gamma_{kl}^i X^k(x) e_j^l \frac{\partial}{\partial e_j^i}. \tag{2.1}$$

For each $m = 1, \cdots, d$ there is a vector field $L_m \in \mathfrak{X}(\mathrm{O}(M))$ such that $L_m(r)$ is the horizontal lift of e_m, where $r = (x, e_1, \cdots, e_d)$. L_m is given by (2.1) with $X^i := e_m^i$. $(L_1, \cdots, L_d)$ is the system of *canonical horizontal vector fields*. The horizontal lift $\tilde{X}$ given in (2.1) of an arbitrary vector field X can be expressed as (see [3, Lemma 2.2])

$$\tilde{X} = \alpha_X^j(r) L_j \tag{2.2}$$

where

$$\alpha_X^i(r) = [e^{-1}]_j^i X^i(x) \tag{2.3}$$

and (e^{-1}) denotes the inverse of the matrix (e_j^i). It is easily shown that $\alpha_X^1, \cdots, a_X^d$ are intrinsic functions on $\mathrm{O}(M)$, i.e., independent of choice of local coordinates.

Let $(\Omega, \mathfrak{F}, (\mathfrak{F}_t), \mathbb{P}, w)$ be the canonical d-dimensional Wiener space and consider the Stratonovich stochastic differential equation (SDE) on $\mathrm{O}(M)$

$$\begin{aligned} &df(r_t) = L_j f(r_t) \circ dw_t^j, \qquad f \in C^\infty(M), \\ &r_0 = r \in \mathrm{O}(M). \end{aligned} \tag{2.4}$$

This equation has a unique solution that defines a diffusion process $r(t)$ evolving in $\mathrm{O}(M)$. By writing (2.4) in terms of the Itô integral, we find that the generator of $r(t)$ is *Bochner's horizontal Laplacian* $\frac{1}{2}\sum_j L_j^2$. Define $x(t, r, w) = \pi r(t)$ (r is the starting point). It is not hard to show that $x(t, r, w)$ has the following property:

$$x(t, T_A r, w) = x(t, r, Aw) \tag{2.5}$$

where A is any orthogonal $d \times d$ matrix, $r = (x, e)$, and $T_A r = (x, \bar{e}_1, \cdots, \bar{e}_d)$ with $\bar{e}_k = A_k^j e_j$. However, d-dimensional Brownian motion is rotation invariant, so Aw is another Brownian motion and (2.5) shows that the law of $x(t, r, w)$ for $r = (x, e)$ depends only on x, not on e. Hence $x(t, r, w)$ is a diffusion process on M. Its generator is the Laplace–Beltrami operator on M:

$$\Delta_M = a^{ij} \frac{\partial^2 f}{\partial x^i \partial x^j} - a^{ij} \Gamma_{ij}^k \frac{\partial f}{\partial x^k}. \tag{2.6}$$

For this reason $x(t, r, w)$ is known as *Brownian motion on M.* This construction is due to Eells and Elworthy [6].

3. Problem formulation. Let M be as above and U be a compact metric space. Consider a family of second-order differential operators given in local coordinates by

$$A^y = \frac{1}{2} a^{ij}(x) \frac{\partial}{\partial x^i} \frac{\partial}{\partial x^j} + b^i(x, y) \frac{\partial}{\partial x^i} \qquad x \in M, \quad y \in U.$$

We assume that a^{ij} is positive definite: $a^{ij}(x)\xi_i\xi_j > 0$ for $\xi \in \mathbb{R}^d$, $\xi \neq 0$, and that $a^{ij} \in C^2(M)$, $b^i \in C^{1,0}(M \times U)$, $i, j = 1, \cdots, d$. We denote by $\mathfrak{U}$ the set of *feedback controls*

$$\mathfrak{U} = \{u: [0, 1] \times M \to U, u \text{ Borel measurable}\}.$$

For each $u \in \mathfrak{U}$ we have a time varying family of differential operators, denoted by slight abuse of notation A^u, defined by

$$A^u = \frac{1}{2} a^{ij}(x) \frac{\partial}{\partial x^i} \frac{\partial}{\partial x^j} + b^i(x, u(t, x)) \frac{\partial}{\partial x^i}. \tag{3.1}$$

We want to regard A^u as the differential generator of a diffusion process (x_t^u) on M over the time interval $[0, 1]$ and then to pose the optimal control problem of finding $u^* \in \mathfrak{U}$ such that

$$\mathbb{E}[\theta(x_1^{u^*})] = \min_{u \in \mathfrak{U}} \mathbb{E}[\theta(x_1^u)] \tag{3.2}$$

where $\theta \in C^2(M)$ is a given function and $\mathbb{E}$ denotes expected value. We could have more general forms of cost function, but we stick to the terminal cost (3.2) for simplicity.

By considering how the coefficients a^{ij} behave under coordinate transformations we can show that the inverse matrix $a_{ij} = [a^{ij}]^{-1}$, which is also symmetric and positive definite, defines a Riemannian metric on M. We can now construct Brownian motion on M as outlined in § 2, i.e., by solving the SDE

$$df(r_t) = L_j f(r_t) \circ dw_t^j, \qquad f \in C^\infty(\mathrm{O}(M)) \tag{3.3}$$

where L_j are the canonical horizontal vector fields and w_t is Brownian motion in $\mathbb{R}^d$. Then $r_t = (x_t, e_t)$ is an $\mathrm{O}(M)$-valued process whose "downstairs" component $x_t = \pi r_t$ is an M-valued diffusion with generator (2.6). To obtain the required generator (3.1) we use the Girsanov transformation. Let

$$\bar{b}^k(x, y) = b^k(x, y) + \tfrac{1}{2} a^{ij}(x) \Gamma_{ij}^k(x), \qquad x \in M, \quad y \in U. \tag{3.4}$$

For each $y \in U$, $\bar{b}^k(\cdot, y)$ defines a vector field X_0^y on M (i.e., the $\bar{b}^k$ transform correctly under coordinate changes). The horizontal lift L_0^y of X_0^y is then given as in (2.3) by

$$L_0^y = \alpha^j(r, y) L_j \tag{3.5}$$

where

$$\alpha^j(r, y) = [e^{-1}]_i^j \bar{b}^i(x, y).$$

Now take $u \in \mathfrak{U}$ and define a probability measure $\mathbb{P}^u$ on $\mathfrak{F}$ by

$$\frac{d\mathbb{P}^u}{d\mathbb{P}} = \exp\left(\sum_{i=1}^d \left(\int_0^1 \alpha^i(r_t, u(t, x_t))\, dw_t^i - \frac{1}{2} \int_0^1 (\alpha^i(r_t, u(t, x_t)))^2\, dt \right) \right\}$$

where (r_t) is the solution of (3.3). Under measure $\mathbb{P}^u$, the process w_t^u given by

$$dw_t^{u,i} = dw_t^i - \alpha^i(r_t, u(t, x_t))\, dt$$

is Brownian motion, and we can rewrite (3.3) as

$$df(r_t) = L_0^u f(r_t)\, dt + L_i f(r_t) \circ dw_t^{u,i}, \qquad f \in C^\infty(\mathrm{O}(M)) \tag{3.6}$$

where

$$L_0^u := \alpha^j(r, u(t,x))L_j.$$

The generator of $r(t)$ is $\tilde{A}^u = L_0^u + \frac{1}{2}\sum_1^d L_i^2$. By taking $f \in C^\infty(M)$ and defining $\tilde{f} = f \circ \pi$ we find that $\tilde{A}^u \tilde{f} = A^u f$, i.e.,

$$M_t^f := f(x_t) - \int_0^t A^u f(s, x_s)\, ds$$

is a martingale on $(\Omega, (\mathfrak{F}_t), \mathbb{P}^u)$, where A^u is given by (3.1), and hence (x_t) is, under $\mathbb{P}^u$, a diffusion process with extended generator A^u, in accordance with the "martingale problem" formulation of diffusion theory [14], [16]. The *cost* associated with $u \in \mathfrak{U}$ is now defined as $J(0, x, u)$ where

$$J(s, x, u) = \mathbb{E}_{s,x}^u[\theta(x_1)] \tag{3.7}$$

and $\mathbb{E}_{s,x}^u$ denotes expectation with respect to the measure $\mathbb{P}_{s,x}^u$ on path space $C([s, 1], O(M))$ generated by the solution of (3.6) with initial condition $r_s = (x, e)$ (e arbitrary). We now have the precisely defined optimal control problem of choosing $u \in \mathfrak{U}$ so as to minimize $J(0, x, u)$.

For technical reasons it is necessary to introduce another class of controls, the so-called *nonanticipative controls.* An *admissible system* is a collection

$$\nu = \{\Xi, \mathfrak{G}, (\mathfrak{G}_t)_{t\in[0,1]}, \mu, (\beta_t, u_t)_{t\in[0,1]}\} \tag{3.8}$$

where $(\Xi, \mathfrak{G}, (\mathfrak{G}_t), \mu)$ is a filtered probability space and $\beta_t(\xi) u_t(\xi)$, $\xi \in \Xi$ are $\mathfrak{G}_t$-predictable processes taking values in $\mathbb{R}^d$, U, respectively, such that (β_t) is a $\mathfrak{G}_t$-Brownian motion. Denote by $\mathfrak{N}$ the set of admissible systems; we call (u_t) a *nonanticipative control* if it is the control process of some $\nu \in \mathfrak{N}$. For any admissible system $\nu \in \mathfrak{N}$ the SDE

$$df(r_t) = \check{L}_0^u f(r_t)\, dt + L_i f(r_t) \circ d\beta_t^i \tag{3.9}$$

has a unique strong (Itô) solution starting at $r_s = (x, e)$, where

$$\check{L}_0^u := \alpha^j(r, u_t(\xi))L_j. \tag{3.10}$$

(We will consistently use the "hacek" notation $\check{L}$ to denote a vector field depending on a control which enters as a "random parameter.") Now define

$$K(s, x, \nu) = \int_\Xi \theta(x_1(\xi))\mu(d\xi)$$

and note that again this does not depend on the initial frame e, since property (2.5) still holds. Of course, (r_t) and (x_t) are not generally Markov processes. We have the inclusion $\mathfrak{U} \subset \mathfrak{N}$ in that each $u \in \mathfrak{U}$ defines the admissible system $\hat{\nu}_u = \{\Omega, \mathfrak{F}, (\mathfrak{F}_t), \mathbb{P}^u, (w_t^u, u_t = u(t, x_t))\}$, and $K(s, x, \hat{\nu}_u) = J(s, x, u)$. Thus

$$\inf_{\nu\in\mathfrak{N}} K(s, x, \nu) \leqq \inf_{u\in\mathfrak{U}} J(s, x, u).$$

4. Dynamic programming. The control problem as formulated in the previous section is to minimize $J(0, x, u)$ given by (3.7) over the class $\mathfrak{U}$ of feedback controls subject to the dynamic constraint (3.6). The Bellman equation for this problem, to be solved for a function $V: [0, 1] \times M \to \mathbb{R}$, is

$$\begin{aligned} &\frac{\partial V}{\partial t}(t, x) + \min_{y\in U}[A^y V(t, x)] = 0, \qquad (t, x) \in [0, 1) \times M, \\ &V(1, x) = \theta(x), \qquad x \in M. \end{aligned} \tag{4.1}$$

THEOREM 4.1. *Suppose that* $a^{ij}(x) \in C^1(M)$, $b^i(x, u) \in C^{1,1}(M \times U)$ *and* $\theta \in C^2(M)$. *Then equation* (4.1) *has a unique solution in* $C^{1,2}([0, 1] \times M)$. *There is Borel function* $u^* : [0, 1] \times M \to U$ *such that*

$$\min_{y \in U} [A^y V(t, x)] = A^{u^*} V(t, x).$$

The proof of this result is given in Appendix A. It is proved by a patching technique: we use the corresponding result for diffusions in bounded regions of $\mathbb{R}^d$ (Theorem VI.6.1 of Fleming and Rishel [8]) in regions covered by a single coordinate chart, and obtain compatability of boundary conditions by a fixed-point argument.

THEOREM 4.2. *Let* u^*, V *be as in Theorem* 4.1. *Then* $u^* \in \mathfrak{U}$ *is optimal in the class* $\mathfrak{N}$ *of nonanticipative controls, and*

$$V(s, x) = \min_{u \in \mathfrak{U}} J(s, x, u) = \min_{\nu \in \mathfrak{N}} K(s, x, \nu).$$

Thus V *is the value function for the control problem.*

Proof. This is the conventional "verification theorem" (cf. Theorem VI.4.1 of [8]). Take any $\nu \in \mathfrak{N}$ as in (3.8) and let $r_t = (x_t, e_t)$ be the corresponding solution of (3.9). Define $\tilde{V} : [s, 1] \times \mathrm{O}(M) \to \mathbb{R}$ by $\tilde{V}(t, r) = V(t, \pi r)$. Then from (3.9) applying the Itô rule and using the fact that $V \in C^{1,2}(M)$ we find that for $t \in [s, 1]$,

$$V(t, x_t) - V(s, x) = \int_s^t \left(\frac{\partial V}{\partial t}(\tau, x_\tau) + A^u V(\tau, x_\tau) \right) d\tau + \int_s^t L_i \tilde{V}(\tau, r_\tau)\, d\beta_\tau^i.$$

From (4.1) the integrand in the first term on the right is nonnegative, and the second term is a martingale (M is compact and the integrand is bounded). Thus the process $t \mapsto V(t, x_t)$ is a μ-submartingale, and hence

$$V(s, x) \leqq \mathbb{E}^{\mu}_{s,x}[V(1, x_1)] = \mathbb{E}^{\mu}_{s,x}[\theta(x_1)] = K(s, x, \nu) \tag{4.2}$$

(we have used the boundary condition from (4.1)). When $\nu = \hat{\nu}_{u^*}$, inequality is replaced by equality in (4.2). The result follows. □

The Bellman equation (4.1) implies that the optimal control u^* satisfies the minimum principle

$$L_0^{u^*} V(t, x) = \min_{y \in U} L_0^y V(t, x) \tag{4.3}$$

where L_0^y is given by (3.5), or, in local coordinates, from (3.4),

$$b^i(x, u^*(t, x)) \frac{\partial V}{\partial x^i}(t, x) = \min_{y \in U} b^i(x, y) \frac{\partial V}{\partial x^i}(t, x).$$

Denote by $\langle \alpha \,|\, X \rangle_x$ the pairing between a 1-form $\alpha \in T_x^*(M)$ and a tangent vector $X \in T_x(M)$, and recall that for $f \in C^1(M)$, df is the 1-form defined by $\langle df \,|\, X \rangle_x = Xf(x)$. We can then write (4.3) as

$$\langle dV(t) \,|\, L_0^{u^*} \rangle_x = \min_{y \in U} \langle dV(t) \,|\, L_0^y \rangle_x.$$

This shows that the "adjoint variable" is the 1-form valued process p_t defined by

$$p_t = dV(t, x_t), \tag{4.4}$$

in that the optimal control u^* minimizes the Hamiltonian functional $\langle p_t \,|\, L_0^y \rangle_{x_t}$ at each t almost surely when p_t is given by (4.4). To characterize p_t further, we need to introduce the derivative or linearized system equations. These are discussed in the next section, and we then return to a characterization of p_t in § 6.

5. Derivative systems of SDEs. We will first discuss derivative systems in the context of SDEs evolving on M. For $X_0, \cdots, X_d \in \mathfrak{X}(M)$ the SDE

$$df(x_t) = X_0 f(x_t)\, dt + X_i f(x_t) \circ dw_t^i, \qquad f \in C^\infty(M) \tag{5.1}$$

has an almost surely unique solution $x_t(x)$ and, furthermore, almost surely the map $x \mapsto x_t(x)$ is a diffeomorphism for each $t \geqq 0$ (see [11], [14]). The Itô version of (5.1) is

$$df(x_t) = \left[X_0 f(x_t) + \frac{1}{2} \sum_{i=1}^{d} X_i^2 f(x_t) \right] dt + X_i f(x_t)\, dw_t^i,$$

which shows that x_t is a diffusion process with generator $X_0 + \frac{1}{2}\sum_i X_i^2$. This is a second-order operator that in local coordinates is similar to (3.1). Let $X \in \mathfrak{X}(M)$ and denote by $\phi_t(x)$ the integral curve of X, i.e., the solution of the ordinary differential equation

$$\frac{d}{dt} f(\phi_t(x)) = Xf(\phi_t(x)), \qquad f \in C^\infty(M),$$

$$\phi_0(x) = x.$$

The *derivative vector field* δX is a vector field on $\mathrm{GL}(M)$ defined by

$$(\delta X) f(r) = \frac{d}{dt} f(\phi_t(x), \phi_{t*} e)|_{t=0}, \qquad r = (x, e) \in \mathrm{GL}(M), \quad f \in C^\infty(\mathrm{GL}(M)).$$

Here $\phi_{t*} e = [\phi_{t*} e_1, \cdots, \phi_{t*} e_d]$ and ϕ_{t*} denotes the derivative map

$$(\phi_{t*} v) g(x) = v(g \circ \phi_t)(\phi_t^{-1}(x)), \qquad g \in C^\infty(M).$$

Now consider the following SDE on $\mathrm{GL}(M)$:

$$\begin{aligned} &df(r_t) = \delta X_0 f(r_t)\, dt + \delta X_i f(r_t) \circ dw_t^i, \qquad f \in C^\infty(\mathrm{GL}(M)), \\ &r_0 = r = (x, e) \in \mathrm{GL}(M). \end{aligned} \tag{5.2}$$

Again, this has an almost surely unique solution, and we can show that

$$r_t(r) = (x_t(x), x_{t*}(e)). \tag{5.3}$$

Taking $(x, v) \in TM$, i.e., $v \in T_x(M)$, we can also think of (5.2) as defining the flow $(x, v) \mapsto (x_t(x), x_{t*}(v))$ on the tangent bundle. When $M = \mathbb{R}^d$, (5.2) is equivalent to the usual system of "linearized equations" written in matrix form. In § 6 below we need to consider the evolution of functionals of the form $\langle W(t, x_t) \,|\, v_t \rangle$ where $W(t, x)$ is a time-varying 1-form field and (x_t, v_t) denotes the solution of (5.2) thought of as a flow on the tangent bundle TM. The following lemma will be useful.

LEMMA 5.1. *Let $\bar{\omega}(x)$ be a 1-form field defined in a neighborhood of $x \in M$, and take $v \in T_x(M)$ and $X \in \mathfrak{X}(M)$. Then*

$$\delta X \langle \bar{\omega}(x) | v \rangle = \langle \nabla_X \bar{\omega} \,|\, v \rangle + \langle \bar{\omega} \,|\, \nabla_v X \rangle$$

and

$$\langle \bar{\omega}(x) \,|\, \nabla_v X \rangle = \langle (\nabla X) \bar{\omega} \,|\, v \rangle.$$

Here ∇_v denotes the covariant derivative, and $(\nabla X)\bar{\omega} := \nabla X(\bar{\omega}, \cdot)$, where ∇X denotes the $(1, 1)$ tensor field given in local coordinates by $(\bar{\omega}, v) \mapsto u_j^i \omega_i v^j$ for $\bar{\omega} = \omega_i\, dx^i$, $v = v^j (\partial/\partial x^j)$ with

$$u_j^i = \frac{\partial}{\partial x^j} X^i(x) + \Gamma_{kj}^i X^k(x).$$

These formulas can be checked directly from the definitions (see Elworthy [7, Lemma VII.9C]). By repeated application of Lemma 5.1 we obtain the following.

COROLLARY 5.2. $(\delta X_0+\frac{1}{2}\sum_{i=1}^{d}(\delta X_i)^2)\langle\bar\omega(x)\,|\,v\rangle=\langle B\bar\omega(x)\,|\,v\rangle$, *where the operator* B *on* 1*-forms is defined by*

$$B\bar\omega(x)=\nabla_{X_0}\bar\omega+(\nabla X_0)\bar\omega+\frac{1}{2}\sum_{i=1}^{d}\{\nabla_{X_i}(\nabla_{X_i}\bar\omega)+2(\nabla X_i)(\nabla_{X_i}\bar\omega)+\nabla(\nabla_{X_i}X_i)\bar\omega\}.$$

We now introduce the notion of the "heat equation for 1-forms." Theorem 5.3 below gives the probabilistic representation for the solution of the heat equation (5.4). While this result is not subsequently used in exactly the form stated here, it needs to be understood since the idea behind it underpins all the developments in § 6. We consider the following equation for a time-varying 1-form field $\bar\omega(t,x)$, where $\theta\in C^2(M)$ is a given function:

$$\begin{aligned}&\frac{\partial}{\partial t}\bar\omega(t,x)+B\bar\omega(t,x)=0, \qquad (t,x)\in[0,1[\times M,\\ &\bar\omega(1,x)=d\theta(x), \qquad x\in M.\end{aligned}\tag{5.4}$$

THEOREM 5.3. *Suppose* $\bar\omega(t,x)$ *satisfies* (5.4). *Then*

$$\langle\bar\omega(s,x)\,|\,v\rangle=\mathbb{E}_{s,(x,v)}\langle d\theta(x_1)\,|\,v_1\rangle\tag{5.5}$$

where $r_t=(x_t,v_t)$ *is the solution of* (5.2) *on* TM *for* $t\in[s,1]$ *with* $r_s=(x,v)$.

Proof. Let $\bar\omega(t,x)$ satisfy (5.4) and define a function f on $[0,1]\times TM$ by

$$f(t,r)=\langle\bar\omega(t,x)\,|\,v\rangle, \qquad r=(x,v).$$

Writing (5.2) in Itô form we have

$$\begin{aligned}df(t,r_t)&=\left[\frac{\partial f}{\partial t}+\delta X_0 f+\frac{1}{2}\sum_i(\delta X_i)^2 f\right]dt+\delta X_i f\,dw^i,\\ &=\left\langle\left(\frac{\partial}{\partial t}+B\right)\bar\omega(t,x)\,|\,v_t\right\rangle dt+\delta X_i f\,dw^i.\end{aligned}\tag{5.6}$$

In view of (5.4) the first term vanishes, and the second term is a martingale. Thus

$$\langle\bar\omega(s,x)\,|\,v\rangle=\mathbb{E}_{s,(x,v)}\langle\bar\omega(1,x_1)\,|\,v_1\rangle.$$

Invoking the boundary condition in (5.4), we obtain (5.5).

6. A characterization of the adjoint process. Let us now return to the control problem of §§ 3 and 4. We cannot directly define the derivative system of the SDE (3.6) under the optimal control u^* since u^* is only known to be a Borel function and the "drift" $b^i{}^*(t,x):=b^i(x,u^*(t,x))$ may fail to be differentiable in x. This is where the nonanticipative controls $\mathfrak{N}$ come in. First, we have the following simple lemma.

LEMMA 6.1. (cf. [10, Lemma VI.7.3]). *Suppose* $f,\ g\in C^1(M)$ *are such that* $f(x)\geqq g(x)$ *for all* $x\in M$ *and* $f(\xi)=g(\xi)$ *at some* $\xi\in M$. *Then* $df(\xi)=dg(\xi)$ (*equality in* $T^*_\xi(M)$).

Fix $(s,\xi)\in[0,1[\times M$ and consider the control problem on the time interval $[s,1]$ with initial condition $x_s=\xi$. Let $u^*(s,x)$ be the optimal control and $\hat u_t:=u^*(t,x_t)$ the control process in the corresponding admissible system $\hat\nu$. Then from Theorem 4.2,

$$V(s,\xi)=J(s,\xi,u^*)=K(s,\xi,\hat\nu).$$

The same control process $\hat{u}_t$ used from another initial state x will be possibly suboptimal, so we have

$$V(s, x) \leqq K(s, x, \hat{\nu}), \qquad x \in M.$$

Applying Lemma 6.1 with $f = K(s, \cdot, \hat{\nu})$ and $g = V(s, \cdot)$ we see that

$$dV(s, \xi) = dK(s, \xi, \hat{\nu}). \tag{6.0}$$

We thus need to calculate dK. We will henceforth write $\hat{\mathbb{P}}$, $\hat{\mathbb{E}}$ for $\mathbb{P}^{u^*}$, $\mathbb{E}^{u^*}$ and β^i for w^{i,u^*}. As a general point of notation, we define $\tilde{\theta}(r) = \theta \circ \pi(r)$ and similarly for other functions on M. We then have for $r = (x, e) \in \mathrm{O}(M)$,

$$K(s, x, \nu) = \tilde{K}(s, r, \nu) = \hat{\mathbb{E}}[\tilde{\theta}(r_{1,s}(r))]$$

where $r_{1,s}(r)$ denotes the solution of (3.9) at time 1 with $r_s = r$. Take $v \in T_x(M)$ and let v' be any element of $T_r\mathrm{O}(M)$ such that $T_\pi v' = v$. Then, with $K(s) := K(s, \cdot, \nu)$,

$$\langle dK(s) \,|\, v\rangle_x = \langle d\tilde{K}(s) \,|\, v'\rangle_r = \hat{\mathbb{E}}[(r_{1,s*}v')\tilde{\theta}(r)] = \hat{\mathbb{E}}\langle d\tilde{\theta} \,|\, r_{1,s*}v'\rangle_{r_1}. \tag{6.1}$$

The derivative system of (3.9) is the SDE

$$\begin{aligned} &df(\varphi_t) = \delta \check{L}_0^{\hat{u}} f(\varphi_t)\, dt + \delta L_i f(\varphi_t) \circ d\beta_t^i, \qquad f \in C^\infty(\mathrm{GL}(\mathrm{O}(M))), \\ &\varphi_s = (r, e') \end{aligned} \tag{6.2}$$

evolving in the frame bundle $\mathrm{GL}(\mathrm{O}(M))$ where (r, e') is a frame in $T_r\mathrm{O}(M)$, with solution

$$\varphi_t = (r_{t,s}(r), r_{t,s*}(e')).$$

(Recall that in (6.2) $\hat{u}_t$ appears as a "random parameter," as in (3.10), and our notational convention of using "$\check{}$" to denote a vector field containing a control as random parameter.) System (6.2) also defines a flow $(r, v') \to (r_{t,s}(r), r_{t,s*}(v'))$ in the tangent bundle $T\mathrm{O}(M)$.

Comparing (5.5) with (6.1) we see that the expressions are the same, except that (6.1) involves the $\mathrm{O}(M)$-valued process r_t, with derivative system (6.2) in $\mathrm{GL}(\mathrm{O}(M))$ in place of the M-valued process x_t with derivative system (5.2) in $\mathrm{GL}(M)$. Thus we can get a representation for dK by reformulating Theorem 5.3 on $\mathrm{O}(M)$. The remainder of this paragraph is devoted to showing that there is a characterization for the value function as a solution of a "heat equation" for 1-form fields.

To consider the heat equation for 1-forms the Laplacian of de Rham-Kodaira is introduced (for more detail see [11]). Define by $\Lambda_p(M)$ the totality of all p-forms on M. An inner product is defined on $\Lambda_p(M)$ denoted as $(\omega, \beta)_p$. The operator $\delta : \Lambda_p(M) \to \Lambda_{p-1}(M)$ is defined by

$$(d\omega, \beta)_p = (\omega, \delta\beta)_{p-1}, \qquad \omega \in \Lambda_{p-1}(M), \quad \beta \in \Lambda_p(M)$$

where $d\omega$ is the exterior derivative[3] of the form ω. The de Rham-Kodaira Laplacian $\square : \Lambda_p(M) \to \Lambda_p(M)$ is defined by

$$\square = -(d\delta + \delta d).$$

The action of this operator on a 1-form $\bar{\omega} = \omega_i(x)\, dx^i$ is given in local coordinates by

$$(\square\bar{\omega})_i = (\Delta\bar{\omega})_i + R_i^j\omega_j \tag{6.3}$$

[3] The exterior derivative $d\omega$ of a p-form ω is a $(p+1)$-form defined by $(d\omega)(x) = (\partial/\partial x^i)\omega_{i_1 i_2 \cdots i_p}(x)\, dx^i \wedge dx^{i_1} \wedge dx^{i_2} \cdots \wedge dx^{i_p}$.

where Δ is the Laplace-Beltrami operator [11, p. 270], and

$$R_i^j = a^{kl} R_{kli}^j .$$

Here

$$R_{jkl}^i = \frac{\partial}{\partial x_k}\Gamma_{lj}^i - \frac{\partial}{\partial x_l}\Gamma_{kj}^i + (\Gamma_{lj}^m \Gamma_{km}^i - \Gamma_{kj}^m \Gamma_{lm}^i)$$

are the components of the curvature tensor. Relation (6.3) is a special case of *Weitzenböck's formula* for differential forms [11, p. 286].

For a 1-form field $\bar{\omega}$ on M we define a function $f: TO(M) \to \mathbb{R}$ by

$$f(r, v) = \langle \bar{\omega} \circ \pi(r) \,|\, T_\pi v \rangle, \qquad r \in O(M), \quad v \in T_r O(M), \tag{6.4}$$

and consider the Itô formula (5.6) with $\check{L}_0^{\hat{u}}$ and L_i replacing X_0, X_i, respectively. Now $\check{L}_0^{\hat{u}}$ is the horizontal lift of $\check{X}_0^{\hat{u}}$, so from Lemma 5.1 we have

$$\delta \check{L}_0^{\hat{u}} f(r, v) = \langle (\nabla_{\check{X}_0^{\hat{u}}} + (\nabla \check{X}_0^{\hat{u}}))\bar{\omega} \circ \pi(r) | T_\pi v \rangle \tag{6.5}$$

where due to the form of the function $f(r, v)$ only the "downstairs" parts of the vector fields remain (see [15]). On the other hand L_i is a canonical horizontal vector field, and from Theorem VII.12.D of [4] we have the following result.

LEMMA 6.2. *Suppose $\bar{\omega}(x)$ is a closed 1-form, i.e., $d\bar{\omega} = 0$, and f is given by* (6.4). *Then*

$$\sum_{i=1}^{d} (\delta L_i)^2 f(r, v) = \langle \Box \bar{\omega} \circ \pi(r) \,|\, T_\pi v \rangle. \tag{6.6}$$

From relations (6.5) and (6.6) we get the following corollary.

COROLLARY 6.3. *Suppose $\bar{\omega}(x)$ is a closed 1-form. Then*

$$\left(\delta \check{L}_0^{\hat{u}} + \frac{1}{2} \sum_{i=1}^{d} (\delta L_i)^2 \right) \langle \bar{\omega} \circ \pi(r) \,|\, T_\pi v \rangle = \langle E\bar{\omega}(x) \,|\, v \rangle$$

where the operator E on 1-forms is defined by

$$E\bar{\omega}(x) = \tfrac{1}{2} \Box \bar{\omega}(x) + (\nabla_{\check{X}_0^{\hat{u}}} + (\nabla \check{X}_0^{\hat{u}}))\bar{\omega}(x).$$

Next consider the following lemma that will be used below.

LEMMA 6.4. *For $X \in \mathfrak{X}(M)$, suppose that $\phi(t, x)$ satisfies the following* PDE *for 1-forms*

$$\begin{aligned} &\frac{\partial \phi}{\partial t}(t, x) = \Box \phi(t, x) + \nabla_X \phi(t, x) + (\nabla X)\phi(t, x), \qquad (t, x) \in]0, 1[\times M, \\ &\phi(0, x) = \tilde{\phi} \end{aligned} \tag{6.7}$$

where $\tilde{\phi}$ is a given 1-form. Then $\phi(t, \cdot)$ is a closed form for each $t \leqq 1$ if $\tilde{\phi}$ is closed.

Proof. First we remark that the Laplacian of de Rham-Kodaira $\Box$ satisfies

$$d \Box \phi = \Box d\phi.$$

Indeed, since $d(d(\cdot)) = 0$,

$$d(d\delta + \delta d) = dd\delta + d\delta d + \delta dd = (d\delta + \delta d)d.$$

Next we remark that

$$\nabla_X \phi + \nabla X \phi = d\langle X \,|\, \phi \rangle.$$

Thus,

$$\frac{\partial\, d\phi}{\partial t} = d\frac{\partial \phi}{\partial t} = d\Box\phi + dd\langle V \,|\, \phi\rangle = \Box\, d\phi, \qquad d\phi(0) = 0.$$

Since zero is the unique solution of this equation in 1-forms [11, § V.5], this shows that $d\phi \equiv 0$, i.e., $\phi(t)$ is a closed form.

Now consider the following heat equation for a time varying 1-form field $W(t, x)$. We write $\hat{X}_0$ in place of $X_0^{u^*}$ (i.e., $\hat{X}_0$ is given in local coordinates by (3.4) with $y = u^*(t, x)$):

$$\frac{\partial}{\partial t} W(t,x) + \frac{1}{2}\Box W(t,x) + \nabla_{\hat{X}_0} W(t,x) + (\nabla \hat{X}_0)\, W(t,x) = 0, \qquad (t,x) \in [0,1[\times M, \tag{6.8}$$

$$W(1,x) = d\theta(x). \tag{6.9}$$

THEOREM 6.5. *Equation* (6.8), (6.9) *has a unique solution* $W(t, x)$ *such that* $W \in W^{1,2}([0,1] \times \mathcal{U})$ *for any open set* $\mathcal{U} \subset M$ *covered by a single coordinate chart, and*

$$W(t,x) = dV(t,x).$$

Proof. Equation (6.8) does not have a $C^{1,2}$ solution since $\hat{X}_0$ is not a smooth function of x. But an argument identical to that of Haussmann [9, Thm. 5.5] shows (see [15]) that it has a unique solution that is (on any coordinate chart) in the Sobolev space $W^{1,2}$ (first derivatives in t, and first and second derivatives in x^i, square integrable). The proof of Lemma 6.4 goes through when (6.7) is solved in $W^{1,2}$. Hence we conclude that the solution $W(t, \cdot)$ of (6.8), (6.9) is a closed 1-form for each $t \in [0, 1]$. (Note that $d\theta$ is closed.) In the following argument, Krylov's extended Itô formula for $W^{1,2}$ functions [12, Thm. 2.10.1] replaces the usual Itô formula.

Take arbitrary $\hat{x} \in M$ and $\hat{v} \in T_x(M)$, denote $\hat{r} = (\hat{x}, \hat{v})$, and let $\tilde{\chi} = (\tilde{r}, \tilde{v})$ be any element of $TO(M)$ such that $x = \pi\tilde{r}$ and $\hat{v} = T_\pi \tilde{v}$. For $(t, \chi) \in [0,1] \times TO(M)$, $\chi = (r, v)$, define

$$f(t,\chi) = \langle W(t, \pi r) \,|\, T_\pi v\rangle.$$

Writing (6.2) in Itô form as the equation for a flow $\chi_t = (r_t, v_t)$ on $TO(M)$ starting at $\chi_s = \tilde{\chi}$, and using Corollary 6.3, we have

$$\begin{aligned} df(t,\chi_t) &= \left[\frac{\partial f}{\partial t} + \delta\check{L}_0^u f + \frac{1}{2}\sum_i^d (\delta L_i)^2 f\right] dt + \delta L_i f\, d\beta_t^i \\ &= \left\langle \left(\frac{\partial}{\partial t} + E\right) W(t, \pi r_t) \,\middle|\, T_\pi v_t \right\rangle dt + \delta L_i f d\beta_t^i. \end{aligned}$$

In view of (6.8) the first term vanishes, and the second term is a martingale. Thus, using the boundary condition (6.9) and the properties of the derivative system (6.2) we see that

$$\begin{aligned} \langle W(s,\hat{x}) \,|\, \hat{v}\rangle_{\hat{x}} &= \langle W(s, \pi\tilde{r}) \,|\, T_\pi \tilde{v}\rangle_{\hat{r}} \\ &= \hat{\mathbb{E}}_{s,\tilde{\chi}} \langle W(1, \pi r_1) \,|\, T_\pi v_1\rangle_{r_1} \\ &= \hat{\mathbb{E}}_{s,\tilde{\chi}} \langle d\tilde{\theta}(r_1) \,|\, T_\pi v_1\rangle \\ &= \hat{\mathbb{E}} \langle d\tilde{\theta} \,|\, r_{1,s*}\tilde{v}\rangle_{r_1}. \end{aligned} \tag{6.10}$$

We thus see from (6.1) and (6.10) that

$$\langle dK(s) \,|\, \hat{v}\rangle = \langle W(s,\hat{x}) \,|\, \hat{v}\rangle_{\hat{x}}.$$

Since this holds for arbitrary $(\hat{x}, \hat{v})$, $dK(s) = W(s, \hat{x})$, and from (6.0),

$$dV(s, \hat{x}) = W(s, \hat{x}).$$

This completes the proof. □

We can now summarize our results, and state the stochastic minimum principle as follows.

THEOREM 6.6. *Let $u^* \in \mathcal{U}$ be an optimal control for the problem* (3.6), (3.7). *Then there exists a unique solution $W(t, x)$ to the heat equation* (6.8), (6.9) *as described in Theorem* 6.5. *Let $X_0(x, y) := X_0^y(x)$ denote the vector field defined by* (3.4). *Then for almost all $(t, x) \in [0, 1] \times M$,*

$$\langle W(t, x) \,|\, X_0(x, u^*(t, x)) \rangle = \min_{y \in U} \langle W(t, x) \,|\, X_0(x, y) \rangle. \tag{6.11}$$

If W is expressed in local coordinates as $W(t, x) = \omega_i(t, x)\, dx^i$ then (6.11) *becomes*

$$\omega_i(t, x) b^i(x, u^*(t, x)) = \min_{y \in U} \omega_i(t, x) b^i(x, y)$$

where b^i are the coefficients appearing in (3.1). *The "adjoint process" as defined in* (4.4) *is thus*

$$p_t := W(t, x_t).$$

Appendix A. This Appendix is devoted to the proof of Theorem 4.1, i.e., showing that the Bellman equation (4.1) has a $C^{1,2}$ solution. Throughout, we shall use a specific system of charts on M constructed as follows. (A similar system was used by Clark [2].) We take an atlas $\{(U_i, g_i), i = 1, \cdots, k\}$ of coordinate charts covering M such that for each i, $g_i(U_i) = B_1$ (the ball of radius one in $\mathbb{R}^d$) and such that $\{E_i, i = 1, \cdots, k\}$ also covers M, where $E_i := g_i^{-1}(B_{3/4})$. Set $D_1 := E_1$ and $D_i := E_i \backslash \cup_{j<i} E_j$ for $i > 1$. Then the D_i are disjoint and cover M.

First we need the following lemmas concerning PDEs in $\mathbb{R}^d$.

LEMMA A1. *Fix $i \in \{1, \cdots, k\}$. Then with the atlas described above $g_i(U_i) = B_1 = \cup_j \tilde{D}_j$ where $\tilde{D}_j = g_i(D_j \cap U_i)$. Denote $Q := [0, 1] \times B_1$ and $Q_j := [0, 1] \times \tilde{D}_j$. Consider the following* PDE *to hold in Q*:

$$\psi_s + A(s)\psi + \Lambda(s, y) = 0 \tag{A1}$$

with boundary data

$$\psi(s, y) = \Psi(s, y), \qquad (s, y) \in \partial^* Q = [0, 1) \times \partial B_1,$$

$$\psi(1, y) = \theta(y), \qquad y \in B_1.$$

Here

$$A(s)\psi = \tfrac{1}{2} a^{ij}(y) \psi_{y_i y_j} + b^i(s, y) \psi_{y_i} + c(s, y)\psi.$$

The coefficients of the PDE (A1) *are supposed to satisfy the following conditions*:

(a) *a^{ij}, b^i, c, Λ satisfy a Hölder condition on Q; $a^{ij} = a^{ji}$ and there exists $\gamma > 0$ such that $a^{ij}(s, y) q_i q_j \geqq \gamma |q|^2$ for all $q \in R^d$.*

(b) *$\theta(y)$ is C^2 on B_1 and $\Psi(s, y) = \tilde{\Psi}(s, y)$ for $t \leqq s < 1$, $y \in \partial B_1$, for some function $\tilde{\Psi}(s, y) : Q \to R$ such that $\tilde{\Psi} | Q_j$ is C^2, $j = 1, \cdots, k$.*

Then (A1) *has a unique solution ψ such that $\psi \in C^{1,2}(\bar{Q}')$ for any open set Q' with $\bar{Q}' \subset Q$.*

Remark. This is the result of Fleming and Rishel [8, Appendix E, first paragraph, p. 208] except that here the boundary data Ψ is "piecewise C^2" as opposed to

everywhere C^2 in [8]. The strategy of the proof is to smooth out the boundary data by convolution of $\tilde{\Psi}$ with a mollifying function, apply the result of [8, Appendix E] and then use a limiting argument. The details, which are lengthy but standard, are omitted here. A complete proof is given in [15].

COROLLARY A2. *Let Q, Ψ be as in Lemma* A1 *and consider data satisfying the following conditions*:

(a) *U is compact*;

(b) *$a^{ij}(x)$ is of class $C^1(\bar{Q})$, and $b^i(x)$ is of class $C^{1,1}(\bar{Q}\times U)$ and the generator A^u is nondegenerate.*

Then the following PDE (*dynamic programming equation*) *in* $\mathbb{R}^d$

$$\frac{\partial V(t,x)}{\partial t}+\min_{y\in U} A^y V=0$$

with boundary data

$$V(s,y)=\Psi(s,y), \qquad (s,y)\in \partial^* Q,$$
$$V(1,y)=\theta(y), \qquad y\in B_1$$

has a unique solution in Q' such that V is in $C^{1,2}(\bar{Q}')$, $\bar{Q}'\subset Q$, where Q' is any open subset of Q.

Proof. This result is the same as Theorem VI.6.1 of Fleming and Rishel [8] except for the conditions on the boundary data Ψ, which is now only piecewise C^2. The proof given in [8, pp. 208–209] goes through unchanged, since Lemma A1 above shows that the statements of [8, first Paragraph, p. 208] are still valid under our wider conditions.

Proof of Theorem 4.1. Let the sets U_i, D_i, $i=1,\cdots,k$ be as defined at the beginning of this Appendix. As is customary we will not distinguish notationally between $U_i\subset M$ and $g_i(U_i)\subset\mathbb{R}^d$. Fix i and in U_i consider the following PDE:

$$\frac{\partial V^i(t,x)}{\partial t}+\min_{y\in U} A^y V^i(t,x)=0, \qquad (t,x)\in[0,1[\times U_i,$$

(A2)
$$V^i(t,x)=\phi(t,x), \qquad 0\leqq t\leqq 1, \quad x\in\partial U_i,$$
$$V^i(1,x)=\theta(x), \qquad x\in U_i$$

where $\phi(t,x)=[0,1]\times M\to\mathbb{R}$ is a given C^2 function. For each (t,x) define

$$L\phi(t,x):=V^i(t,x), \qquad x\in D_i.$$

From results for PDEs in $\mathbb{R}^d$ [5] we know that

$$L\phi(t,x)\in C^{1,2}([0,1]\times D_i)$$

but $L\phi(t,x)$ is "piecewise-continuous" on $([0,1]\times M)$ since in each D_i a "different" PDE is considered. Clearly for every choice of initial point $(t,x)\in[0,1)\times D_i$

$$L\phi(t,x)=\min_{u\in\mathrm{II}} \mathbb{E}^u_{t,x}[\theta(x_1)I_{1=\sigma_i}+\phi(\sigma_i x_{\sigma_i})I_{1>\sigma_i}]$$

where I_A is the characteristic function of A and σ_i denotes the first hitting time of $([0,1]\times\partial U_i)\cup(\{1\}\times U_i)$ by the controlled system of § 3 starting at $x_t=x$. We will define $T_1=\sigma_i$ when $x\in D_i$. We now define $L^2\phi=L(L\phi)$ by replacing ϕ by $L\phi$ in (A2) i.e., solving the PDE (A2) with the following boundary conditions:

$$V^i(t,x)=L\phi(t,x), \qquad 0\leqq t\leqq 1, \quad x\in\partial U_i,$$
$$V^i(t,x)=\theta(x), \qquad x\in U_i.$$

These are the boundary conditions considered in Lemma A1. Then for such $(t, x) \in [0, 1) \times D_i$

$$L^2\phi(t, x) := L(L\phi)(t, x) = V^i(t, x)$$

$$= \min_{u \in \mathrm{ll}} \mathbb{E}^u_{t,x} \left\{ \theta(x_1) I_{1=\sigma_i} + \sum_{j \neq i} L\phi(\sigma_i, x_{\sigma_i}) I_{1>\sigma_i} I_{x_{\sigma_i} \in D_j} \right\}$$

from which is follows that for $x \in D_i$

$$L^2\phi(t, x) = \min_{u \in \mathrm{u}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{1=T_2} + \phi(T_2, x_{T_2}) I_{1>T_2}]$$

where T_2 is defined by

$$T_2 := 1 \wedge \inf\{t > \sigma_i : x_t \in \partial U_{i(x_{\sigma_i})}\}$$

and $i(\cdot)$ is the indicator function $i(\xi) = j$ if $\xi \in D_j$. From Corollary A2 it follows that $L^2\phi(t, x)$ is in $C^{1,2}([0, 1) \times D_i)$, but again "piecewise-continuous" on $[0, 1] \times M$. Similarly, for each (t, x)

$$L^n\phi(t, x) = \min_{u \in \mathrm{ll}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{1=T_n} + \phi(T_n, x_{T_n}) I_{1>T_n}]$$

where T_n is an exit time defined by

$$T_n := 1 \wedge \inf[t > T_{n-1} : x_t \in \partial U_{i(x_{T_{n-1}})}].$$

Now $V^n(t, x) \equiv L^n\phi(t, x)$ is the value of a control problem stopped at T_n. The sequence of exit times T_n is strictly increasing by the construction of the disjoint sets D_j and the nondegeneracy of our problem. It has been proved by Clark [2] that $\sup_n T_n = \infty$ so by the above construction $L^n\phi(t, x)$ is defined for all n. We will show below that as $n \to \infty$, $V^n(t, x)$ converges to the value function of the unstopped problem (4.1). The space $C([0, 1] \times M)$ is a Banach space with the norm

$$\|\phi\| = \sup_{(x,t) \in M \times [0,1]} |\phi(t, x)|.$$

Next we shall show that the map L is a contraction, and hence that $\|L^n\phi - V\| \to 0$, where V is a fixed point of L. Indeed, in D_i

$$L\phi(t, x) - L\psi(t, x) = \min_{u \in \mathrm{ll}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{1=\sigma_i} + \phi(\sigma_i, x_{\sigma_i}) I_{I>\sigma_i}]$$

$$- \min_{u \in \mathrm{ll}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{1=\sigma_i} + \psi(\sigma_i, x_{\sigma_i}) I_{1>\sigma_i}]$$

$$\leqq \mathbb{E}^{u^*}[\theta(x_1) I_{1=\sigma_i}) + \phi(\sigma_i, x_{\sigma_i}) I_{1>\sigma_i}]$$

$$- \mathbb{E}^{u^*}[\theta(x_1) I_{1=\sigma_i} + \psi(\sigma_i, x_{\sigma_i}) I_{1>\sigma_i}]$$

where u^* is the control that minimizes

$$\mathbb{E}^u_{t,x}[\theta(x_1) I_{1=\sigma_i} + \psi(\sigma_i, x_{\sigma_i}) I_{1>\sigma_i}].$$

This can be constructed by a selection theorem as in Fleming and Rishel (see [8, Lemma VI.6.1]). It follows that

$$L\phi - L\psi \leqq \|\phi - \psi\| \mathbb{P}^{u^*}_{t,x}[\sigma_i < 1].$$

Interchanging the roles of ϕ and ψ we conclude that

$$|L\sigma(t, x) - L\psi(t, x)| \leqq \|\phi - \psi\| \mathbb{P}^{u^*}_{t,x}[\sigma_i < 1], \qquad x \in D_i,$$

and hence that $\|L\phi - L_\psi\| \leqq \rho \|\phi - \psi\|$, where $\rho = \max_i \rho_i$ and

$$\rho_i = \sup_{x \in D_i} \mathbb{P}^{u^*}_{t,x}[\sigma_i < 1].$$

Since (D_i) is a finite partition, the contraction property is established once we show that $\rho_i < 1$ for each i. First, since $D_i \subset \text{int}\{U_i\}$, it is a standard result for nondegenerate diffusions that

$$\mathbb{P}^{u^*}_{t,x}[\sigma_i = 1] > 0 \quad \text{for each } x \in D_i,\ t \in [0, 1].$$

Next we remark that $\psi(t, x) = \mathbb{P}^{u^*}_{t,x}[\sigma_i = 1]$ is the solution of the following PDE on U_i:

$$\frac{\partial \psi}{\partial t} + A^{u^*}\psi = 0,$$

$$\psi(1, x) = 1, \qquad x \in U_i,$$

$$\psi(t, x) = 0, \qquad (t, x) \in [0, 1[\times \partial U_i.$$

Hence $\psi(t, x)$ is continuous and by the above argument $\psi(t, x) > 0$ for each x. Thus on the compact set $\bar{D}_i \subset U_i$

$$\pi_i := \min_{x \in D_i} \mathbb{P}^{u^*}_{t,x}[\sigma_i = 1] > 0$$

and thus

$$\rho_i = \sup_{x \in D_i} \mathbb{P}^{u^*}_{t,x}[\sigma_i < 1] < 1.$$

This shows that L is a contraction mapping and hence has a unique fixed point V. We now show that the function V is independent of the construction of the disjoint sets D_i and the defined sequence of stopping times T_i. Indeed, the probabilistic formula for $L^n\phi$ gives

$$L^n\phi = \min_{u \in \mathcal{U}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{1 = T_n} + \phi(T_n, x_{T_n}) I_{T_n < 1}].$$

We observe that as $n \to \infty$ we have $T_n \to \infty$ almost surely and

$$L^n\phi \to \min_{u \in \mathcal{U}} \mathbb{E}^u_{t,x}[\theta(x_1)] \equiv V$$

since

$$\sup_{u \in \mathcal{U}} |\mathbb{E}^u_{t,x}[(\theta(x_1) + \phi(T_n, x_{T_n})) I_{T_n < 1}]| \leqq \|\theta + \phi\| \sup_{u \in \mathcal{U}} \mathbb{P}^u_{t,x}[T_n < 1] \to 0 \quad \text{as } n \to \infty.$$

The above argument essentially means that the probability to pay a penalty $\phi(T_n, x_{T_n})$ at T_n as $n \to \infty$ is negligible. Therefore the fixed point of the map L is equal to the value function V of the unstopped problem and does not depend on the construction of the disjoint set D_i.

For any $x \in D_1$ we know by a standard dynamic programming argument that V satisfies

$$V(t, x) = \min_{u \in \mathcal{U}} \mathbb{E}^u_{t,x}[\theta(x_1) I_{\tau = 1} + V(t, x_\tau) I_{\tau < 1}]$$

where $\tau := 1 \wedge \inf\{s \geqq t: x_s \in \partial D_1\}$. This is a control problem on a single coordinate chart, and we know from results of [8] that $V(t, x) = \tilde{V}(t, x)$ where $\tilde{V}(t, x)$ is the unique $C^{1,2}([0, 1[\times D_1)$ solution of the Bellman equation

$$\frac{\partial \tilde{V}}{\partial t}(t, x) + \min_{y \in U}[A^y \tilde{V}(t, x)] = 0, \qquad (t, x) \in [0, 1[\times D_1,$$

$$\tilde{V}(t, x) = V(t, x), \qquad (t, x) \in [0, 1[\times \partial D_1,$$

$$\tilde{V}(t, x) = \theta(x), \qquad x \in D_1.$$

Thus V is $C^{1,2}$ at any $[t, x] \in [0, 1[\times \text{int}\{D_1\}$. Since D_1 was arbitrary, this shows that in fact $V \in C^{1,2}([0, 1] \times M)$.

REFERENCES

[1] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[2] J. M. C. CLARK, *An introduction to stochastic differential equations on manifolds*, in Geometric Methods in System Theory, D. Q. Mayne and R. W. Brockett, eds., Reidel, Boston, MA, 1974.

[3] M. H. A. DAVIS AND M. P. SPATHOPOULOS, *Pathwise nonlinear filtering for nondegenerate diffusions with noise correlation*, SIAM J. Control Optim., 25 (1987), pp. 260–278.

[4] M. H. A. DAVIS AND P. P. VARAIYA, *Dynamic programming conditions for partially-observable stochastic systems*, SIAM J. Control Optim., 11 (1973), pp. 226-261.

[5] T. E. DUNCAN, *Dynamic programming optimality criteria for stochastic system in Riemannian manifolds*, Appl. Math. Optim., 3 (1978), pp. 191–208.

[6] J. EELLS AND K. D. ELWORTHY, *Stochastic dynamical systems*, in Control Theory and Topics in Functional Analysis III, International Atomic Energy Agency, Vienna, 1976, pp. 179–185.

[7] K. D. ELSWORTHY, *Stochastic Differential Equations on Manifolds*, Cambridge University Press, Cambridge, 1982.

[8] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, New York, 1975.

[9] U. G. HAUSSMANN, *On the adjoint process for optimal control of diffusion processes*, SIAM J. Control Optim., 19 (1981), pp. 221–234.

[10] ———, *A Stochastic Maximum Principle for Optimal Control of Diffusions*, Pitman Research Notes in Mathematics 151, Longman, London, 1987.

[11] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam; Kodansha, Tokyo, 1981.

[12] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, Berlin, New York, 1980.

[13] H. J. KUSHNER, *On the stochastic maximum principle: fixed time of control*, J. Math. Anal. Appl., 11 (1965), pp. 78–92.

[14] L. C. G. ROGERS AND D. WILLIAMS, *Diffusions, Markov Processes and Martingales*, Vol. II, John Wiley, Chichester, 1987.

[15] M. P. SPATHOPOULOS, *Filtering and stochastic control for diffusions on manifolds*, Ph.D. thesis, Imperial College of Science and Technology, London, 1986.

[16] D. W. STROOCK, *Lectures on Topics in Stochastic Differential Equations*, Tata Institute Series No. 68, Narosa, New Delhi; Springer-Verlag, Berlin, New York, 1982.

[17] R. B. VINTER, *New results on the relationship between dynamic programming and the maximum principle*, Math. Control Signals Sys., 1 (1988), pp. 97–105.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1108-1135, September 1989

012

STOCHASTIC APPROXIMATION AND LARGE DEVIATIONS: UPPER BOUNDS AND w.p.1 CONVERGENCE*

PAUL DUPUIS† AND HAROLD J. KUSHNER‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. With probability one convergence results are obtained for stochastic recursive approximation algorithms under very general conditions. The gain sequence $\{a_n\}$ can go to zero very slowly and state-dependent noise, discontinuous dynamical equations, and the projected or constrained algorithm are all treated. The basic technique is the theory of large deviations. Prior results obtained via this theory are extended in many directions. Let $\dot{x} = \bar{b}(x)$ denote the "mean" equation for the algorithm, let $\delta > 0$ be given, and let $G(\theta)$ be a neighborhood of a stable point θ of that ordinary differential equation. Then, asymptotic upper bounds to $a_N \log P\{X_n \notin G(\theta),\ n \geqq N \,|\, |X_N - \theta| \leqq \delta\}$ are obtained. These are often more informative than the usual classical rate of convergence results (that use a "local linearization") and, furthermore, are obtained for the constrained and nonsmooth cases, for which there are no "rate of convergence" results. The methods are also used to extend currently available upper bounds for algorithms with constant gains, with simpler proofs.

Key words. stochastic approximation, large deviations, recursive algorithms, errors for tracking systems

AMS(MOS) subject classifications. 60F10, 62L20, 93E10, 93E12

1. Introduction. We obtain with probability one (w.p.1) convergence results as well as useful (nonclassical) estimates of "rate of convergence" for fairly general stochastic approximation (SA) processes such as (1.1), via the theory of large deviations (R^r = Euclidean r-space)

$$X_{n+1} = X_n + a_n b_n(X_n, \xi_n), \quad X_n \in R^r, \quad 0 < a_n \to 0, \quad \Sigma a_n = \infty, \quad n \geqq 0. \tag{1.1}$$

We also treat the projection algorithm (1.2), where π_G denotes the nearest point of a compact and not necessarily convex set G:

$$X_{n+1} = \pi_G(X_n + a_n b_n(X_n, \xi_n)). \tag{1.2}$$

Such algorithms have been the subject of considerable attention [1]-[4], [8], [28], [29], [31], under a great variety of conditions. They appear in various guises in many places in control and communication theory.

In (1.1), the $\{\xi_n\}$ is a random process that might be state dependent itself in the sense that $P\{\xi_{n+1} \in A \,|\, \xi_i, X_i, i \leqq n\} \neq P\{\xi_{n+1} \in A \,|\, \xi_i, i \leqq n\}$. The b_n might simply be a function of X_n, ξ_n. The formulation allows $\{b_n\}$ to be a sequence of vector-valued (R^r) mutually independent, but not necessarily stationary, *random fields* parametrized by X_n, ξ_n. In this case b_n is characterized by the distribution function (which will depend on n in the nonstationary case)

$$P\{b_n \in B \,|\, X_i, \xi_i, b_{i-1}, i \leqq n\} = P\{b_n \in B \,|\, X_n, \xi_n\}. \tag{1.3}$$

The actual model used includes these as special cases and is defined at the beginning of § 2. There are many applications where the random field notation is useful since it

* Received by the editors May 27, 1988; accepted for publication (in revised form) December 13, 1988.

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003. The research of this author was supported in part by Office of Naval Research contract N-00014-83-K-0542 and National Science Foundation grant DMS-8511470.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by Air Force Office of Scientific Research contract AFOSR-85-0315, National Science Foundation grant ECS-8505674, and Army Research Office contract GC-A-662998.

is awkward or difficult to express explicitly all the random variables that might be involved (e.g., the X_n, ξ_n might determine other random variables that are used, in turn, to calculate X_{n+1} from X_n). For example, consider an adaptive routing problem, where X_n denotes the routing parameter and ξ_n the (vector) buffer occupancies at time n. Then b_n might be a random variable that depends on, e.g., arrivals, completed services, and on acceptances of arrivals, at time n, and each of these might be related to X_n, ξ_n only statistically—but the exact relation is either too complicated to write (perhaps involving a sum of indicator functions of various possible events) or not necessary to write.

If b_n is simply a function of X_n, ξ_n, $(b_n = b(X_n, \xi_n))$, then we call it a *deterministic random field.* Even in this case, the ξ_n might be state dependent, correlated, or $b(\cdot)$ might be discontinuous. If $\{b_n\}$ is a deterministic random field, we write it simply as $b(X_n, \xi_n)$. Of course, since $\{\xi_n\}$ is a random sequence, $\{b(X_n, \xi_n)\}$ is not deterministic, in the usual sense.

Perhaps the weak convergence-based methods [3], [5], [6] are the most powerful general methods for dealing with the asymptotic properties of (1.1) or (1.2). The conditions for the validity of such methods are often readily verifiable. One common approach is to derive an ordinary differential equation (ODE) for the "mean" dynamics $\dot{x} = \bar{b}(x) = Eb(x, \xi)$ (where this is well defined) and to show that the asymptotic path of $\{X_n\}$ is arbitrarily close to that of the asymptotic solutions to $\dot{x} = \bar{b}(x)$ in the sense of the weak convergence theory. Typically, under some stability property of the ODE, this method locates the points (or point) near which $\{X_n\}$ spends "nearly all of its time." Nevertheless, there is still considerable interest in actual w.p.1 convergence. A powerful method would use a weak convergence approach to find the "asymptotic" points or sets, and then use a "local" method to show w.p.1 convergence of $\{X_n\}$ to an appropriate stable point of the ODE, under the usual condition that some compact set in its domain of attraction is entered infinitely often (that would itself often be shown by a weak convergence-based method).

Among methods that can be used to prove w.p.1 convergence, those based on the theory of large deviations have a number of advantages. The methods developed here yield a fairly unified approach for problems with state-dependent noise and discontinuous dynamics as well as for constrained problems. These facts imply the availability of a rather powerful technique for getting w.p.1 convergence. The state-dependent noise model used here is more general than allowed in [3] and [7], and is essentially the same as that in [28], [29], and [31]. The mathematical development here seems to be no more complicated than the powerful "martingale" based methods of [4], [8], [28], and [29]. Particularly in view of the fact that once certain basic general results are proved, we do not need to treat the various special cases independently. We can handle more slowly (and erratically) converging gains, the constrained case, the random field model, and get a very informative estimate of the rate of convergence even when the classical "local" smoothness conditions are violated. This latter point is particularly important. Although we do not do so here, the basic theorems can also be used to obtain results for models with random gains. The underlying idea is relatively simple. The proof of w.p.1 convergence reduces to showing the differentiability of a certain function defined in terms of conditional expectations taken with respect to the process itself. We show this readily for the usual cases covering the bulk of work in the literature, as well as for some new cases. New cases appearing in future applications can be dealt with in the same way. Our methods can have difficulty with problems where the qth moments of the ξ_n or $b_n(x, \xi_n)$ grow too fast as $q \to \infty$ (say, faster than those for b_n = Gaussian), but this rarely seems to be a serious problem in applications.

Typically, the large deviations estimates involve both an upper and a lower bound for a (suitably normalized) probability of a "rare" event (say the event that the stochastic approximation (asymptotically) escapes from a small neighborhood of a stable point of $\dot{x} = \bar{b}(x)$). To get the w.p.1 convergence here, only an *upper bound* is needed, and this allows a result under weaker conditions than would be required if both bounds were desired. The upper bound serves as a useful indicator of the rate of convergence, perhaps even more useful than that obtained by the classical methods. It is obtainable for constrained problems and is often obtainable even for nonstationary problems, in contrast to the classical "rate" results.

The "rate" calculated by the classical methods is just the asymptotic variance of $(X_n - \theta)/a_n^{1/2}$, where θ is the limit point. Its derivation requires a certain "regularity" in the way $a_n \to 0$, and a local expansion of the dynamics about θ. Assuming appropriate smoothness (usually twice differentiability of $b(x, \xi)$ at $x = \theta$, which is not needed by the large deviations method) of b for x near θ, the classical rate depends only on the gradient of $Eb(x, \xi)$ for $x = \theta$ and on the statistics of $\{b(\theta, \xi_n)\}$. In many applications, we are more interested in an (suitably normalized) estimate of the probability that the path $\{X_n, \infty > n \geqq N\}$ will escape from some given neighborhood of θ for large N. This would involve the full stabilizing effect of the dynamics and "destabilizing" effect of the noise in that interval, and such a useful estimate is obtainable from our results.

Our rate estimate takes the following form. Let D denote a compact set in the domain of attraction of a stable point θ of the ODE $\dot{x} = \bar{b}(x)$ and with $\theta \in D^0$, the interior of D. Let $\delta > 0$ be given. Let $A_D(T)$ denote the set of continuous functions $\phi(\cdot)$ with $|\phi(0) - \theta| \leqq \delta$ and $\phi(t) \notin D$ for some $t < T$. We will exhibit a function $\bar{L}(\phi, \dot{\phi}, t) \geqq 0$ that is zero if and only if $\dot{\phi} \equiv \bar{b}(\phi)$ and a function $\bar{S}(x, T, \phi)$:

$$\bar{S}(x, T, \phi) = \begin{cases} \displaystyle\int_0^T \bar{L}(\phi(s), \dot{\phi}(s), s)\, ds & \text{(for } \phi \text{ absolutely continuous, with } \phi(0) = x\text{)}, \\ \infty & \text{(otherwise)}, \end{cases}$$

such that

$$\overline{\lim_{n\to\infty}}\, a_n \log P\{X_m \notin D, \text{ some } m \geqq n \,|\, |X_n - \theta| \leqq \delta\} \leqq - \inf_{\substack{\phi \in A_D(T) \\ T > 0}} \bar{S}(\phi(0), T, \phi) < 0. \tag{1.4}$$

The right-hand side of (1.4) can yield estimates that are very useful for a "rate" of convergence, and for the dependence of this rate on the behavior of the algorithm in the set of interest D, as well as for the comparison of algorithms.

In [9]-[11], sharp upper and lower bounds have been obtained for SA algorithms by the methods of large deviations theory, and a great deal of useful information has been presented concerning the bounds and the structure of the H and L-functionals. These references require that $a_n \to 0$ in special ways, the noise is "exogenous," and the dynamical term b is a smooth function of x. The methods are unable to handle the constrained problems. Strictly speaking, the results in these references are not w.p.1 convergence results. They deal with sequences of sequences $\{X_m^n, m \geqq 0\}$, $n = 1, 2, \cdots$, defined by $X_{m+1}^n = X_m^n + a_{n+m} b(X_m^n, \xi_{n+m})$, $X_0^n = x$. Although the analysis of such processes is basic to the convergence result, in this paper we deal with the actual process itself. Also, since we are concerned with upper (large deviations) bounds only, we use $\overline{\lim}$ to define the various functionals, rather than lim, as illustrated in the sequel. This allows a result under weaker conditions on the $\{a_n, \xi_n, b_n\}$, as will be seen below.

To obtain the general results, we proceed as follows. First, a general and somewhat abstract assumption (Assumption 2.1 in § 2) is made. Under this (and some more

readily verifiable and reasonable) assumptions, the w.p.1 convergence theorem and the rate of convergence estimates are obtained (§ 3). In § 4, we show that all parts of Assumption 2.1 (except for part (i)) hold for the case of bounded dynamical terms. The extension to the unbounded case is disposed of in § 7. Then, to get the large deviation upper bound (and hence the w.p.1 result) for specific examples, we need only verify that a certain $\bar{H}(x, \alpha, t)$-functional, defined in § 4, has an α-derivative at $\alpha = 0$. This is done in § 5 for models covering the bulk of cases dealt with in the literature, and a general method for verification is discussed. (It is worth noting that the α-differentiability requirement is in a certain sense necessary and sufficient for a stochastic process to possess a nontrivial large deviation upper bound. The sufficiency is the point of Theorems 4.1 and 7.1, while the necessity can be shown to follow (under weak assumptions on the process) from the application of a theorem of Varadhan's [21] to the problem of evaluating (4.1).) Section 6 treats the constrained case, basically by showing how a "continuity" theorem for large deviations estimates can be used to carry the "unconstrained" result over to the constrained case. There are numerous advantages to our method, in comparison with existing methods. A more detailed comparison is given in § 8.

The key to the entire development is the result of § 4, obtaining an upper bound in a very quick and efficient way, without using the usually complicated sequence of estimates for special cases often associated with large deviations bounds. This method applies equally well to the case where the gains $\{a_n\}$ are constant. For this case, we obtain large deviations upper bounds for stochastic difference equations of the type in [27], but with weaker conditions and an easier proof, and the same "action functional." See, in particular, § 5.1.a and Example 7.1.

2. Background and assumptions. In this section, we introduce some rather detailed assumptions that will be used to prove the main convergence theorem and the rate estimate in § 3. The basic large deviation upper bound given in Assumption 2.1(iv) is of course not simple to verify, but is used simply to facilitate the proofs in § 3. We give some examples of processes that satisfy these assumptions at the end of this section as well as in §§ 5 and 7, where we give readily verifiable sufficient conditions that cover a wide variety of applications.

Assume that our algorithm is given by

$$X_{n+1} = X_n + a_n F_n. \tag{2.1}$$

Define

$$t_n = \sum_{0}^{n-1} a_i, \qquad t_0 = 0. \tag{2.2}$$

We define a continuous parameter interpolated version of $\{X_n\}$ by

$$X(t) = [(t - t_n)X_{n+1} + (t_{n+1} - t)X_n]/a_n, \qquad t \in [t_n, t_{n+1}], \tag{2.3}$$

and the interpolated version *starting at time* t_N by

$$X^N(t) = X(t + t_N). \tag{2.4}$$

When we say that $\overline{\lim}_n f_n \leqq f$ uniformly in a parameter α for some sequence $\{f_n\}$ and some f, we mean that for each $\varepsilon > 0$, there is $n_\varepsilon < \infty$ such that $\sup_{n \geqq n_\varepsilon} f_n \leqq f + \varepsilon$ for all α. For a set or point A, we use $N_h(A)$ for the h-neighborhood of A. Let $C[0, T]$ denote the set of R^r-valued functions on $[0, T]$ with the sup-norm topology.

ASSUMPTION 2.1. There exists a family of σ-algebras $\mathcal{F}_n \supset \sigma(X_i, i \leqq n)$, and a functional $\bar{S}(x, T, \phi)$ defined for $x \in R^r$, $T > 0$, and $\phi(\cdot) \in C[0, T]$ with the following properties:

(i) There exists $\bar{b}(x)$ such that $\bar{S}(x, T, \phi) = 0$ if and only if $\dot{\phi} = \bar{b}(\phi)$ almost surely, $\phi(0) = x$;

(ii) $\bar{S}(x, T, \phi) \geqq 0$;

(iii) Given compact $F \subset R^r$, $T > 0$, and $s \in [0, \infty)$, the set $\{\phi: \phi(0) \in F, \bar{S}(\phi(0), T, \phi) \leqq s\}$ is compact;

(iv) Given compact $F \subset R^r$, $T > 0$, $h > 0$ and $s \in [0, \infty)$, $\overline{\lim}_N a_N \log P\{X^N(\cdot) \notin N_h(\Phi(X^N(0), T, s)) \mid \mathcal{F}_N\} \leqq -s$, uniformly in $X^N(0) \in F$, and ω (w.p.1), where $\Phi(x, T, s) = \{\phi: \bar{S}(x, T, \phi) \leqq s\}$.

Remarks. (1) By part (iii) above, the functional $\bar{S}(\cdot, T, \cdot)$ is lower semicontinuous (l.s.c.).

(2) We consider $X^N(\cdot)$ in part (iv) as restricted to $[0, T]$.

(3) A weaker form of (iv) is actually sufficient for our needs below. If we assume that given $M_1 < \infty$ we can prove the lim sup is uniform save on a set of ω's with probability less than $\exp -M_1/a_N$, then the conclusions of § 3 may still be obtained, since the sets where the uniformity fails are negligible from the point of view of the large deviations estimates. We refer to this extended assumption as Assumption 2.1^e. Let us mention an example where these considerations are useful. If F_n takes the form $b(X_n) + \sigma(X_n)\xi_n$, where the ξ_n are obtained as the solution of a stable linear system driven by independently and identically distributed (i.i.d.) and zero mean Gaussian noise, then for $b(\cdot)$ and $\sigma(\cdot)$ Lipschitz we obtain Assumption 2.1(iv) only with the weaker uniformity described above, but this is enough to obtain the w.p.1 convergence. We will return to this example at the end of this section.

(4) In §§ 4, 5, and 7 we prove Assumption 2.1 for many interesting cases, and indicate the connection between the statistics of the process $\{X_n\}$ and the functions $\bar{b}(x)$ and $\bar{S}(x, T, \phi)$. We note at this time that Assumption 2.1 implies that the conditional distribution of $X^N(\cdot)$ given $X^N(0) = x$ converges weakly to the measure concentrated at the point $\phi(\cdot)$, where $\dot{\phi} = \bar{b}(\phi)$, $\phi(0) = x$ (if this solution is unique), and that the probability that $X^N(\cdot)$ deviates from $\phi(\cdot)$ by more than $\gamma > 0$ on any interval $[0, T]$ (in the sup-norm sense) decays as does $\exp -\delta/a_N$, for some $\delta > 0$.

(5) From Assumption 2.1(iv) we may obtain the following [15, Thm. 3.3]. Let compact $F \subset R^r$, $T > 0$, $\delta > 0$ and $s \geqq 0$ be given. Then there is $N_0 < \infty$ such that for any $x \in F$, any closed set $A \subset C[0, T]$ satisfying $\inf_{\phi \in A} \bar{S}(x, T, \phi) \geqq s$, and any $N \geqq N_0$, we have

$$a_N \log P\{X^N(\cdot) \in A \mid \mathcal{F}_N, X^N(0) = x\} \leqq -s + \delta \quad \text{a.s.} \tag{2.5}$$

(6) The uniformity of the estimates in Assumption 2.1(iv) with respect to ω imply that (2.5) continues to hold if we replace N by any stopping time $M \geqq N_0$ almost surely.

(7) The t-dependence of $\bar{S}$ occurs owing to the fact that a_n are not constant. See, e.g., Example 2.1, and the results in § 5.

The Limit ODE. To get any sort of useful convergence for $\{X_n\}$, the ODE

$$\dot{x} = \bar{b}(x) \tag{2.6}$$

must have at least one stable point. We assume the following.

ASSUMPTION 2.2. The ODE (2.6) has a unique solution for each initial condition and there is a point θ that is asymptotically (not necessarily globally) stable in the sense of Lyapunov, with domain of attraction Λ.

Assumption 2.2 implies that for any compact $G \subset \Lambda$ and $\delta > 0$ there is $T < \infty$ such that all solutions originating in G are in $N_\delta(\theta)$, a δ-neighborhood of θ, for $t \geqq T$.

Finally, we state the slowest rate at which we can allow $a_n \to 0$.

ASSUMPTION 2.3. For every $\delta > 0$, $\sum_n \exp -\delta/a_n < \infty$, and $\sum a_n = \infty$.

For example, let $a_n = c_n/\log n$, with $c_n \to 0$ and $\sum a_n = \infty$. Then Assumption 2.3 holds. If c_n does not go to zero, then there will not be convergence w.p.1. This is the case in the "annealing" process [30].

Example 2.1. We take $F_n = b(X_n) + \sigma(X_n)\xi_n$ in (2.1), where $b(\cdot)$ and $\sigma(\cdot)$ are Lipschitz and $a_n = 1/n^\gamma$, $0 < \gamma \leqq 1$. We assume that $\xi_{n+1} = A\xi_n + B\theta_n$, with $\{\theta_n\}$ an i.i.d. mean-zero Gaussian sequence, and that the roots of A are contained in the interior of the unit circle. Then [11] Assumption 2.1^e holds, with

$$\bar{S}(x, T, \phi) = \begin{cases} \int_0^T \bar{L}(\phi, \dot{\phi}, s)\, ds & \text{if } \phi \text{ is absolutely continuous and } \phi(0) = x, \\ \infty & \text{otherwise,} \end{cases}$$

$$\bar{L}(x, \beta, s) = (\beta - b(x))'[(\sigma(x)(A-I)^{-1}B)(\sigma(x)(A-I)^{-1}B)']^{-1}(\beta - b(x))h(s)/2,$$

(if the indicated inverse exists) and $\bar{b}(x) = b(x)$. The function $h(s)$ is $\exp s$ if $\gamma = 1$, and 1 if $\gamma < 1$. If the indicated inverse does not exist, $\bar{L}$ takes a different form [11], although Assumption 2.1^e continues to hold with $\bar{b}(x) = b(x)$. For additional examples see [11].

The sequences $\{a_n\}$ and the functions $b(\cdot)$ and $\sigma(\cdot)$ considered above are more "regular" than is actually needed to obtain Assumption 2.1^e. This is because the indicated reference was concerned with *both upper and lower large deviation bounds.* For the upper bound alone, which is the only part of the theory we use, much less "regularity" is required. We shall see more of this in §§ 5 and 7. In particular, Example 7.2 contains Example 2.1 as a special case.

3. The basic convergence theorem.

THEOREM 3.1. *Suppose that Assumptions* 2.1, 2.2, *and* 2.3 *hold, and that there is a compact neighborhood* $G(\theta)$ *of* θ, *with* $G(\theta) \subset \Lambda^0$, *and such that there is* (*almost surely*) *a* (*random*) *sequence* $\{n_i\}$ *satisfying* $X_{n_i} \in G(\theta)$. *Then* $X_n \to \theta$ w.p.1.

Remarks. If not all paths visit some neighborhood of θ infinitely often (i.o.) then we will have $X_n \to \theta$ w.p.1 with respect to those paths that do. It is expected that the recurrence condition would be verified by a weak convergence argument.

Proof. For $\delta > 0$, let $N_\delta(\theta)$ denote $\{x: |\theta - x| \leqq \delta\}$. We will first prove that if $\{X_n\}$ visits $G(\theta)$ infinitely often w.p.1, then $\{X_n\}$ visits $N_\delta(\theta)$ infinitely often w.p.1. We can suppose that $N_\delta(\theta) \subset G(\theta)$.

Owing to the stability assumption (Assumption 2.2), there is $T_1 < \infty$ such that if $\phi(\cdot)$ satisfies $\dot{\phi} = \bar{b}(\phi)$ and $\phi(0) = x \in G(\theta)$, then $\phi(t) \in N_{\delta/2}(\theta)$ for $t \geqq T_1$. Let $\Delta > 0$, and set $T_1' = T_1 + \Delta$. Define the set of paths

$$A_1 = \{\phi(\cdot) \in C[0, T_1']: \phi(0) \in G(\theta), \phi(t) \notin N_\delta(\theta) \quad \text{for some } t \in [T_1, T_1']\}.$$

We claim that there is a $c_1 > 0$ such that ($\bar{A}$ and A^0 denote the closure and interior, respectively, of the set A)

$$\inf_{\phi \in \bar{A}_1} \bar{S}(\phi(0), T_1', \phi) = c_1 > 0. \tag{3.1}$$

If (3.1) does not hold, then there is a sequence $\{\phi_i(\cdot)\} \subset A_1$ such that $\bar{S}(\phi_i(0), T_1', \phi_i) \to 0$. By Assumption 2.1(iii), the set $\{\phi_i(0)\}$ is precompact. We extract a subsequence (again indexed by i) such that $\{\phi_i(\cdot)\}$ converges, and denote the limit by $\phi^*(\cdot)$. The lower semicontinuity of $\bar{S}(\phi(0), T_1', \phi)$ implies that $\bar{S}(\phi^*(0), T_1', \phi^*) = 0$, which

implies that

$$\dot{\phi}^* = \bar{b}(\phi^*), \qquad \phi^*(0) \in G(\theta).$$

Since by its definition $\bar{A}_1$ does not contain a solution of $\dot{\phi} = \bar{b}(\phi)$, we obtain a contradiction.

Define the events

$$E_n^1 = \{X_n \in G(\theta), X(t+t_n) \notin N_\delta(\theta) \text{ for some } t \in [T_1, T_1']\}.$$

We have $E_n^1 = \{X^n(\cdot) \in A_1\}$. Assumptions 2.1 and 2.3, and (3.1) then imply that

$$\sum_n P\{E_n^1\} \leqq \sum_n P\{E_n^1 \mid X_n \in G(\theta)\} < \infty$$

and the Borel-Cantelli Lemma gives

$$P\{E_n^1 \text{ occurs infinitely often}\} = 0.$$

Define $m(t) = \min\{n: t_n \leqq t\}$. Since $a_n \to 0$, we conclude that the event $\{X_n \in G(\theta), X_{m(T_1'+t_n)} \notin N_\delta(\theta)\}$ occurs only finitely often, w.p.1. Thus, if $\{X_n\}$ visits $G(\theta)$ infinitely often, w.p.1, it must visit $N_\delta(\theta)$ infinitely often w.p.1 for each $\delta > 0$.

Next let $\delta_1 > 0$ be such that $N_{\delta_1}(\theta) \subset \Lambda^0$. By the stability assumption there is $\delta_1 > \delta_2 > 0$ such that for any $x \in N_{\delta_2}(\theta)$, if $\phi(\cdot)$ satisfies $\dot{\phi} = \bar{b}(\phi)$ and $\phi(0) = x$, then $\phi(t) \in N_{\delta_1/2}(\theta)$ for $t \geqq 0$. By the preceding argument, we can assume that

$$P\{X_n \in N_{\delta_2}(\theta) \text{ infinitely often}\} = 1.$$

Let $T_2 < \infty$ be such that if $\phi(\cdot)$ satisfies $\dot{\phi} = \bar{b}(\phi)$ and $\phi(0) = x \in N_{\delta_2}(\theta)$, then $\phi(t) \in N_{\delta_2/2}(\theta)$ for all $t \geqq T_2$. Define

$$A_2 = \{\phi(\cdot) \in C[0, T_2]: \phi(0) \in N_{\delta_2}(\theta) \text{ and there is } t \leqq T_2 \text{ such that } \phi(t) \notin N_{\delta_1}(\theta) \text{ and/or } \phi(T_2) \notin N_{\delta_2}(\theta)\}.$$

By an argument analogous to that used to get (3.1), there is a $c_2 > 0$ such that

$$\inf_{\phi \in \bar{A}_2} \bar{S}(\phi(0), T_2, \phi) = c_2 > 0.$$

Define $E_n^2 = \{X_n \in N_{\delta_2}(\theta)$ and there is $t \leqq T_2$ such that $X(t+t_n) \notin N_{\delta_1}(\theta)$ and/or $X(T_2+t_n) \notin N_{\delta_2}(\theta)\}$. Then $E_n^2 = \{X^n(\cdot) \in A_2\}$.

Let $\{m_k\}$ denote the return times of $\{X_n\}$ to $N_{\delta_2}(\theta)$ and note that the $\delta_i > 0$ can be made arbitrarily small. Since $m_k < \infty$ for all k w.p.1, to prove the theorem it is sufficient to show that $\lim P\{X_{m_k+i} \notin N_{\delta_1}(\theta), \text{ for some } i < \infty\} = 0$. But this holds if we show that (w.p.1)

$$\lim_n P\{X_{n+i} \notin N_{\delta_1}(\theta), \text{ for some } i < \infty \mid X_n \in N_{\delta_2}(\theta)\} = 0. \tag{3.2}$$

We have the obvious inclusion (letting A^c denote the complement of the set A)

$$\begin{aligned}
&\{X_{n+i} \notin N_{\delta_1}(\theta), \text{ for some } i < \infty \text{ and } X_n \in N_{\delta_2}(\theta)\} \\
&\quad \subset \{X_i \notin N_{\delta_1}(\theta) \text{ for some } m(jT_2+t_n) < i \leqq m(jT_2+T_2+t_n) \\
&\qquad \text{and/or } X_{m(jT_2+T_2+t_n)} \notin N_{\delta_2}(\theta), \text{ for some } 0 \leqq j < \infty, \text{ and } X_n \in N_{\delta_2}(\theta)\} \\
&\quad = \bigcup_{0 \leqq j < \infty} E^2_{m(jT_2+t_n)} \bigcap_{i<j} (E^2_{m(iT_2+t_n)})^c \cap \{X_n \in N_{\delta_2}(\theta)\}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&P\{X_{n+i} \notin N_{\delta_1}(\theta), \text{ for some } i < \infty \mid X_n \in N_{\delta_2}(\theta)\} \\
&\qquad \leqq \sum_{j=0}^{\infty} P\left\{ E^2_{m(jT_2+t_n)} \,\middle|\, \bigcap_{i<j} (E^2_{m(iT_2+t_n)})^c \cap \{X_n \in N_{\delta_2}(\theta)\} \right\}.
\end{aligned} \tag{3.3}$$

Note that for any fixed j, inclusion in the conditioning set in the jth term in the sum implies that $X_{m(jT_2+t_n)} \in N_{\delta_2}(\theta)$. Thus (3.2) follows from (3.3), and Assumptions 2.1 and 2.3. □

To obtain our "rate" estimate, we make a weak assumption on the form of $\bar{S}(x, T, \phi)$.

ASSUMPTION 3.1. There exists a measurable function $\bar{L}(x, \beta, s) \geqq 0$ such that

$$\bar{S}(x, T, \phi) = \begin{cases} \int_0^T \bar{L}(\phi, \dot{\phi}, s)\, ds & \text{if } \phi \text{ is absolutely continuous and } \phi(0) = x, \\ \infty & \text{otherwise} \end{cases}$$

and that as a function of t, $\bar{L}(x, \beta, t)$ is nondecreasing for each pair (x, β).

This representation of $\bar{S}(x, T, \phi)$ holds for every example of a "decreasing gain" stochastic approximation process having a large deviation upper bound known to the authors. For the processes studied in this paper, we will define L by (4.1) and (4.2) below.

THEOREM 3.2. *Assume Assumptions* 2.1, 2.2, 2.3, *and* 3.1 *hold, and, in addition, that given* $\varepsilon > 0$ *there is* $\bar{N} < \infty$ *such that* $a_i / a_N \leqq 1 + \varepsilon$ *for all* $i \geqq N \geqq \bar{N}$. *Then, for* $G(\theta)$ *a neighborhood of* θ

$$\begin{aligned} &\overline{\lim_N}\, a_N \log P\{X_n \notin G(\theta), \text{ for some } n \geqq N \,|\, |X_N - \theta| \leqq \delta\} \\ &\leqq - \inf_{\substack{\phi : |\phi(0) - \theta| \leqq \delta \\ \phi(t) \notin G(\theta), \text{some } t < \infty}} \bar{S}(\phi(0), t, \phi) \equiv -\bar{S}^* \end{aligned} \tag{3.4}$$

uniformly in ω, w.p.1.

Remark. The stability assumption and the lower semicontinuity of $\bar{S}(\cdot, T, \cdot)$ imply that the right side of (3.4) is strictly negative for small enough $\delta > 0$.

Proof. Let $\bar{T} > 0$ be fixed and $T \geqq \bar{T}$, and let $N_\delta(\theta) \subset G(\theta)$. Define that set of paths

$$\begin{aligned} A(T) = \{\phi : |\phi(0) - \theta| \leqq \delta, \text{ and either } \phi(t) \notin G(\theta) \text{ for some } t \leqq T, \\ \text{or } |\phi(t) - \theta| \geqq \delta/2 \text{ for } \bar{T} \leqq t \leqq T\}. \end{aligned}$$

We claim that for large enough T,

$$\inf_{\phi \in \overline{A(T)}} \bar{S}(\phi(0), T, \phi) \geqq \bar{S}^*. \tag{3.5}$$

First note that the same proof as that of (3.1) implies there are $c_3 > 0$ and $T_3 < \infty$ such that if we define $A_3 = \{\phi : \phi(0) \in G(\theta), \phi(T_3) \notin N_{\delta/2}(\theta)\}$, then

$$\inf_{\phi \in \bar{A}_3} \bar{S}(\phi(0), T_3, \phi) \geqq c_3.$$

Let i equal the integer part of $(T - \bar{T})/T_3$. Then for the paths in $A(T)$ that do not escape from $G(\theta)$ and for which $|\phi(t) - \theta| \geqq \delta/2$ for $\bar{T} \leqq t \leqq T$, we have

$$\bar{S}(\phi(0), T, \phi) \geqq i c_3,$$

which implies (3.5) (when T is large).

Now define the stopping times τ_i^N by $\tau_0^N = N$, $\tau_{i+1}^N = \inf\{n \geqq m(t_{\tau_i^N} + \bar{T}) : X_n \notin N_\delta(\theta)\} \wedge \inf\{n \geqq \tau_i^N : X_n \notin G(\theta)\}$ and the events

$$E_i^N = \{X_{\tau_{i+1}^N} \notin G(\theta) \text{ or } t_{\tau_{i+1}^N} - t_{\tau_i^N} \geqq T\} \subset \{X^{\tau_i^N}(\cdot) \in A(T)\}.$$

We use the following estimate that is derived in the same way as (3.3).

$$P\{X_n \notin G(\theta), \text{ for some } n \geqq N \mid X_N \in N_\delta(\theta)\}$$

$$\leqq \sum_{j=0}^{\infty} P\left\{ E_j^N \,\middle|\, \bigcap_{i<j} (E_i^N)^c \cap \{X_N \in N_\delta(\theta)\} \right\} \quad \text{(w.p.1)}. \tag{3.6}$$

Fix $h_1 > 0$. By Assumption 2.1 and (3.5), an upper bound (w.p.1) to the right-hand side of (3.6) is given by (for any $h_2 \geqq 0$)

$$\sum_{i=N}^{\infty} \exp -(\bar{S}^* - h_1)/a_i = (\exp -(\bar{S}^* - h_2)/a_N) \sum_{i=N}^{\infty} \exp [-(\bar{S}^* - h_1)/a_i + (\bar{S}^* - h_2)/a_N]$$

when N is large. Thus (3.4) follows if we prove that given $h_2 > 0$ there is $h_1 > 0$, $\bar{N} < \infty$, and $M < \infty$ so that for $N \geqq \bar{N}$,

$$\sum_{i=N}^{\infty} \exp [-(\bar{S}^* - h_1)/a_i + (\bar{S}^* - h_2)/a_N] \leqq M. \tag{3.7}$$

To prove (3.7), take $\varepsilon = (h_2/8\bar{S}^*) \wedge \frac{1}{2}$, and $h_1 = \varepsilon \bar{S}^*/2$. Pick $\bar{N}$ large enough so that $a_i/a_N \leqq 1+\varepsilon$ for $i \geqq N \geqq \bar{N}$. Then for each i such that $a_i/a_N \geqq 1-\varepsilon$ we have

$$[-\bar{S}^* + h_1 + (\bar{S}^* - h_2)a_i/a_N]/a_i \leqq [h_1 + \varepsilon\bar{S}^* - h_2(1-\varepsilon)]/a_i \leqq [-h_2/4]/a_i.$$

On the other hand, if $a_i/a_N < 1-\varepsilon$, we obtain the following bound for the exponent:

$$[h_1 - \varepsilon\bar{S}^*]/a_i = [-\varepsilon\bar{S}^*/2]/a_i.$$

Hence (3.7) follows from Assumption 2.3. □

4. A proof showing that Assumption 2.1(ii)–(iv) holds for bounded noise. In this section we make the following simplifying assumption.

ASSUMPTION 4.1. The sequence $\{|F_n|\}$ is (almost surely) bounded by $K < \infty$.

The main result of this section is that under this condition, the proof that Assumption 2.1 holds reduces to verifying the existence of the derivative at $\alpha = 0$ of the function $\bar{H}(x, \alpha, t)$ defined below. In § 5 we will show the existence of this derivative for a wide variety of processes.

For $\phi \in C[0, T]$ let $D^\Delta\phi(t) = \phi(t+\Delta) - \phi(t)$. We define the function $\bar{H}(x, \alpha, t)$ by

$$\bar{H}(x, \alpha, t) = \overline{\lim_{\Delta \to 0}} \overline{\lim_{N \to \infty}} \overline{\lim_{\substack{y \to x \\ s \to t}}} \operatorname*{ess\,sup}_{\omega} a_N \log E[\exp \langle \alpha, D^\Delta X^N(s) \rangle / a_N \mid \mathscr{F}_N, X^N(s) = y]/\Delta. \tag{4.1}$$

Although the expression defining $\bar{H}$ appears formidable, we shall see in § 5 that it simplifies greatly for "typical" classes of models. The advantage of dealing with the definition as stated is that it yields Theorem 4.1 below under minimal assumptions.

Remark on the use of $\overline{\lim}$ *rather than* lim *in* (4.1). The use of $\overline{\lim}$ is somewhat equivalent to taking a worst case. For example, let $b_n(x, \xi) = b(x) + \xi$, where $\{\xi_n\}$ is a sequence of zero-mean mutually independent Gaussian random variables with covariances $\{\Sigma_n\}$, and let $a_n = 1/n^\gamma$, where $\gamma \in (0, 1)$. Since

$$\frac{1}{n} \log E \exp \left\langle \alpha, \sum_{N+1}^{N+n} (b(x) + \xi_j) \right\rangle = \langle \alpha, b(x) \rangle + \frac{1}{2n} \sum_{N+1}^{N+n} \langle \alpha, \Sigma_j \alpha \rangle,$$

we can prove that the $\overline{\lim}$ in (4.1) is just $\langle \alpha, b(x) \rangle + \alpha' \Sigma \alpha / 2$, where Σ is the $\overline{\lim}$ of $(1/n) \sum_{N+1}^{N+n} \Sigma_i$ in the sense of nonnegative definite matrices. In many problems, the

dynamics are stable enough so that if the noise terms are multiplied by some factor (to take, say, Σ_n to Σ) we still have the required "stability" to get the desired w.p.1 convergence. Additional examples appear in § 5.

Owing to its definition, the function $\bar{H}(x, \alpha, t)$ enjoys the following properties:

(i) The convergence in (4.1) is uniform in ω (almost surely);

(ii) $\bar{H}(x, \alpha, t)$ is convex in α and is upper semicontinuous in (x, t);

(iii) $\bar{H}(x, 0, t) = 0$ for all pairs (x, t);

(iv) $\bar{H}(x, \alpha, t) \leqq K|\alpha|$ for all pairs (x, t), where K is an upper bound to $|F_n|$.

Define $\bar{L}(x, \beta, t)$ as the Legendre transform (in α) of $\bar{H}(x, \alpha, t)$:

$$\bar{L}(x, \beta, t) = \sup_{\alpha} [\langle \alpha, \beta \rangle - \bar{H}(x, \alpha, t)]. \tag{4.2}$$

The properties of $\bar{H}$ mentioned above imply the following properties of $\bar{L}$:

(v) $\bar{L}(x, \beta, t)$ is convex in β and is lower semicontinuous in (x, β, t);

(vi) $\bar{L}(x, \beta, t) \geqq 0$ for all (x, β, t);

(vii) $\bar{L}(x, \beta, t) = +\infty$ if $|\beta| > K$, for all pairs (x, t).

Here (ii)$\Rightarrow$(v), (iii)$\Rightarrow$(vi), (iv)$\Rightarrow$(vii).

Finally, we define an "action functional" in terms of $\bar{L}$ by

$$\bar{S}(x, T, \phi) = \int_0^T \bar{L}(\phi, \dot{\phi}, s)\, ds \tag{4.3}$$

if $\phi(0) = x$ and ϕ is absolutely continuous, and $\bar{S}(x, T, \phi) = \infty$ otherwise.

Remark. We refer to $\bar{S}$ as an action functional, even though it yields only upper bounds. In the sequel we will define an action functional simply by writing an equation like (4.3), the "absolute continuity" and $\phi(0) = x$ qualifications being understood.

THEOREM 4.1. *Assume Assumption* 4.1, *and define* $\bar{H}$, $\bar{L}$, *and* $\bar{S}$ *by* (4.1), (4.2), *and* (4.3), *respectively. Then parts* (ii)-(iv) *of Assumption* 2.1 *hold.*

Remarks. With this theorem, it is clear that as far as bounded noise is concerned, the main task associated with any specific model is verifying Assumption 2.1(i) for a vector field $\bar{b}(\cdot)$ with the correct stability properties. We discuss this problem at length in § 5. The proof below will yield Assumption 2.1(iv) with the same σ-algebra as that used to define $\bar{H}$ in (4.1), which is in fact open to choice. However, we want to choose $\mathcal{F}_N$ to obtain the "best" upper bound (i.e., largest $\bar{L}$, or smallest $\bar{H}$). Usually the "best" choice is quite obvious. We also have flexibility with regard to the gain sequence $\{a_i\}$. If the sequence $\{a_i|F_i|/\bar{a}_i\}$ is bounded (almost surely), then we obviously obtain an upper bound with normalizing sequence $\{\bar{a}_i\}$, although with a possibly different functional $\bar{S}$.

Proof of (ii) *and* (iii). The nonnegativity of $\bar{L}$ implies that of $\bar{S}$. The lower semicontinuity of $\bar{L}$ and the lower bound $\bar{L}(x, \beta, t) = \infty$ for $|\beta| > K$ are sufficient to prove [14] that $\bar{S}(\phi(0), T, \phi)$ is lower semicontinuous in $\phi(\cdot)$. Since $\bar{S}(\phi(0), T, \phi) < \infty$ implies $|\dot{\phi}(t)| \leqq K$ almost surely, Ascoli's Theorem yields the compactness.

(iv) Fix compact $F \subset R^r$, $T > 0$, $s < \infty$, and $h > 0$. Let θ: $[0, T] \to R^r$ be continuous, and for the remainder of the proof let $y(\cdot)$ be a Lipschitz continuous function (with constant K) with $y(0)$ in F.

For $\gamma > 0$, define

$$\bar{H}^{\gamma}(x, \alpha, t) = \sup_{|x-y| \leqq \gamma} \sup_{|t-s| \leqq \gamma} \bar{H}(y, \alpha, s).$$

Then $\bar{H}^{\gamma}$ is convex in α and upper semicontinuous in (x, t), and $\bar{H}^{\gamma}(x, \alpha, t) \downarrow \bar{H}(x, \alpha, t)$ as $\gamma \to 0$. Let $\bar{L}^{\gamma}$ denote the Legendre transform of $\bar{H}^{\gamma}$.

For each N, set

$$G_N^\gamma(\theta, y) = \sum_N^{m(t_N+T)} \langle \theta(t_i - t_N), y(t_{i+1} - t_N) - y(t_i - t_N) \rangle - \sum_N^{m(t_N+T)} \bar{H}^\gamma(y(t_i - t_N), \theta(t_i - t_N), t_i - t_N) a_i. \tag{4.4}$$

Due to the definition of $\bar{H}^\gamma$,

$$\lim_{N\to\infty} G_N^\gamma(\theta, y) \geqq \int_0^T \langle \theta(t), \dot{y}(t) \rangle \, dt - \int_0^T \bar{H}^{2\gamma}(y(t), \theta(t), t) \, dt \equiv G^{2\gamma}(\theta, y)$$

uniformly in $y(\cdot)$. For functions $y(\cdot)$ which are not Lipschitz with constant K, we set $G^{2\gamma}(\theta, y) = +\infty$. Note that $G^{2\gamma}(\theta, y)$ is l.s.c. in $y(\cdot)$ for each $\theta(\cdot)$.

Let $x \in F$, and let $\varepsilon > 0$ be given. By the definition (4.1), for any compact set F' there is $\tau > 0$ such that $\Delta \leqq \tau$, $N \geqq 1/\tau$, $|y - x| \leqq \tau$, and $|t - s| \leqq \tau$ imply that

$$E[\exp \langle \alpha, D^\Delta X^N(s) \rangle / a_N \mid \mathcal{F}_N, X^N(s) = y] \leqq \exp (\bar{H}^\gamma(x, \alpha, t) + \varepsilon)\Delta / a_N$$

(almost surely in ω), whenever $x \in F'$. If we choose F' to contain the range of $X^N(\cdot)$ on the interval $[0, T]$, then

$$a_N \log E\left[\exp \sum_{i=0}^{T/\Delta - 1} (\langle \alpha, X^N(i\Delta + \Delta) - X^N(i\Delta) \rangle - \bar{H}^\gamma(X^N(i\Delta), \alpha, i\Delta)\Delta)/a_N \,\middle|\, \mathcal{F}_N, X_N = x \right] \leqq T\varepsilon.$$

By using the continuity of $\theta(\cdot)$ and the fact that ε can be made arbitrarily small, we get

$$\overline{\lim_{N\to\infty}} a_N \log E[\exp G_N^\gamma(\theta, X^N)/a_N \mid \mathcal{F}_N, X_N = x] \leqq 0 \tag{4.5}$$

uniformly in ω (almost surely) and $x \in F$. Let closed A be given. Then, by Chebyshev's inequality, for each $\theta(\cdot) \in C([0, t]; R^r)$

$$P\{X^N(\cdot) \in A \mid \mathcal{F}_N, X_N = x\} \leqq E\left[\exp (G_N^\gamma(\theta, X^N) - \inf_{\phi \in A} G_N^\gamma(\theta, \phi))/a_N \mid \mathcal{F}_N, X_N = x \right].$$

We also have

$$\overline{\lim_N} \left[-\inf_{\phi \in A} G_N^\gamma(\theta, \phi) \right] \leqq -\inf_{\phi \in A} G^{2\gamma}(\theta, \phi).$$

Combining these facts, we obtain

$$\overline{\lim_N} a_N \log P\{X^N(\cdot) \in A \mid \mathcal{F}_N, X_N = x\} \leqq -\inf_{\phi \in A} G^{2\gamma}(\theta, \phi), \tag{4.6}$$

uniformly in ω almost surely and $x \in F$, for each $\gamma > 0$.

Define (the set $\Phi(x, T, s)$ is defined in Assumption 2.1(iv))

$$s^*(\gamma) = \inf_{\phi \notin N_h(\Phi(x,T,s))} \sup_{\theta \in C([0,T];R^r)} G^\gamma(\theta, \phi).$$

We shall assume that for each $\gamma > 0$ that $s^*(\gamma) < \infty$. The case $s^*(\gamma) = \infty$ requires only obvious modifications. Let $\delta > 0$ be given. For any $\phi(\cdot) \notin N_h(\Phi(x, T, s))$ there is

$\theta(\phi)(\cdot)$ such that $G^\gamma(\theta(\phi), \phi) \geqq s^*(\gamma) - \delta$. The l.s.c. of $G^\gamma(\theta, \cdot)$ implies that there is a neighborhood $N(\phi)$ of $\phi(\cdot)$ (with radius less than $h/2$) such that for all $\psi(\cdot) \in N(\phi)$, $G^\gamma(\theta(\phi), \psi) \geqq s^*(\gamma) - 2\delta$. Choose $\phi_i(\cdot)$, $i = 1, \cdots, I$, such that the $N(\phi_i)$ cover the compact set given by $\{\phi(\cdot): \phi(0) = x, |\dot{\phi}(t)| \leqq K, \ t \leqq T, \ x \in F\}$. Let $I(x, s) = \{i: \phi_i(\cdot) \notin N_h(\Phi(x, T, s))\}$. By (4.6) and the comments above in this paragraph,

$$\begin{aligned}
&\overline{\lim_N} a_N \log P\{X^N(\cdot) \notin N_h(\Phi(x, T, s)) \mid \mathcal{F}_N, X_N = x\} \\
&\quad \leqq \overline{\lim_N} a_N \log \sum_{i \in I(x,s)} P\{X^N(\cdot) \in N(\phi_i) \mid \mathcal{F}_N, X_N = x\} \\
&\quad \leqq -s^*(\gamma) + 2\delta
\end{aligned}$$

(uniformly in $x \in F$ and in ω almost surely). Since the inequality between the second and third lines holds if only one term appeared in the sum, it holds as stated. If we show that $\liminf_{\gamma \to 0} s^*(\gamma) \geqq s$, then (since $\delta > 0$ is arbitrary) we have proved Assumption 2.1(iv).

Now fix $\phi(\cdot)$, and consider $\sup_\theta G^\gamma(\theta, \phi)$. Assume $\int_0^T \bar{L}^\gamma(\phi, \dot{\phi}, t)\, dt < 0$. (The case $\int_0^T \bar{L}^\gamma(\phi, \dot{\phi}, t)\, dt = \infty$ is handled similarly.) Given $n < \infty$, we can find a measurable function $\theta^n(\cdot)$ such that

$$\begin{aligned}
\int_0^T [\langle \theta^n(t), \dot{\phi}(t) \rangle - \bar{H}^\gamma(\phi(t), \theta^n(t), t)]\, dt &\geqq \int_0^T \sup_\alpha [\langle \alpha, \dot{\phi}(t) \rangle - \bar{H}^\gamma(\phi(t), \alpha, t)]\, dt - 1/n \\
&= \int_0^T \bar{L}^\gamma(\phi(t), \dot{\phi}(t), t)\, dt - 1/n.
\end{aligned}$$

Next choose a sequence of continuous functions $\theta_i^n(\cdot)$ such that $\theta_i^n \to \theta^n$ almost everywhere. By dominated convergence

$$\lim_i G^\gamma(\theta_i^n, \phi) \geqq \int_0^T \bar{L}^\gamma(\phi(t), \dot{\phi}(t), t)\, dt - 1/n.$$

Therefore,

$$s^*(\gamma) = \inf_{\phi \notin N_h(\Phi(x,T,s))} \int_0^T \bar{L}^\gamma(\phi(t), \dot{\phi}(t), t)\, dt.$$

By extracting from a minimizing sequence $\phi^\gamma(\cdot)$ a subsequence such that $\dot{\phi}^\gamma(\cdot)$ converges weakly and applying Fatou's lemma, we obtain

$$\liminf_{\gamma \to 0} s^*(\gamma) \geqq \inf_{\phi \notin N_h(\Phi(x,T,s))} \bar{S}(x, T, \phi) \geqq s. \qquad \square$$

Remark. Suppose that in the definition of $\bar{H}$ (given by (4.1)) we restrict the ess sup to $\omega \in \Omega_N$, where $\overline{\lim}_N a_N \log P(\Omega \backslash \Omega_N) = -\infty$. In this case the proof of Theorem 4.1 will yield parts (ii), (iii), and (iv) of Assumption 2.1^e, and not Assumption 2.1. However, as remarked below Assumption 2.1, this is sufficient for the convergence proof.

With this theorem we have proved (for the case of bounded $\{F_n\}$) that for $\bar{H}$, $\bar{L}$, and $\bar{S}$ defined through (4.1), (4.2), and (4.3), respectively, we obtain Assumption 2.1 parts (ii), (iii), and (iv), respectively. The only part of Assumption 2.1 not shown is the existence of $\bar{b}(x)$ such that $\bar{L}(x, \beta, t) = 0$ if and only if $\beta = \bar{b}(x)$. We take this matter up in the next section, where we consider this problem in the context of several specific system models. We show this to be true if and only if $\bar{H}_\alpha(x, 0, t)$ exists, in which case we identify $\bar{b}(x)$ as the "asymptotic mean drift" at the point x.

5. On verifying Assumption 2.1(i) for bounded noise. In this section we consider the problem of verifying the only part of Assumption 2.1 not covered by Theorem 4.1. While the proof of Theorem 4.1 made no use of the properties of the noise term $\{F_n\}$ (aside from boundedness), its conclusions may well be vacuous if we do not prove something similar to Assumption 2.1(i) as well. Suppose, for example, that we have $\bar{H}(x, \alpha, t) = K|\alpha|$ for all (x, t). In this case we obtain

$$\bar{L}(x, \beta, t) = \begin{cases} 0, & |\beta| \leqq K, \\ \infty, & |\beta| > K, \end{cases} \tag{5.1}$$

which implies $\bar{S}(\phi(0), T, \phi) < \infty$ if and only if ϕ is Lipschitz with constant K, in which case it is also true that $\bar{S}(\phi(0), T, \phi) = 0$. But, then Assumption 2.1(iv) is nothing more than the trivial statement that $X^N(\cdot)$ is Lipschitz with constant K w.p.1. For Assumption 2.1(iv) to be truly meaningful, we must at least have $\bar{L}(x, \beta, t) > 0$ for some points (x, β, t) with $|\beta| < K$. The proof of such a fact will clearly depend on the model chosen for $\{F_n\}$. In this section we consider a number of interesting models, and in fact prove that Assumption 2.1(i) holds. As a preliminary to the results on specific models, we present two general theorems that are useful in many cases. The first theorem spells out the relationships between the existence of a derivative (in α) of $\bar{H}(x, \alpha, t)$ at $\alpha = 0$, the existence of a unique β such that $\bar{L}(x, \beta, t) = 0$, and the "mean asymptotic" dynamics of $X^N(\cdot)$, given $X_N = x$.

THEOREM 5.1. *Define $\bar{H}(x, \alpha, t)$ and $\bar{L}(x, \beta, t)$ by* (4.1) *and* (4.2), *and assume that $\bar{H}(x, \alpha, t)$ is differentiable in α at $\alpha = 0$ for all (x, t). Then $\bar{H}_\alpha(x, 0, t)$ is independent of t, and $\bar{L}(x, \beta, t) = 0$ if and only if $\beta = \bar{H}_\alpha(x, 0, t)$. Furthermore,*

$$\bar{H}_\alpha(x, 0, t) = \lim_{\Delta \to 0} \lim_{N \to \infty} \lim_{\substack{y \to x \\ s \to t}} E[D^\Delta X^N(s) \mid \mathcal{F}_N, X^N(s) = y]/\Delta \quad \text{(a.s.)}. \tag{5.2}$$

Remarks. Recall that $D^\Delta x(s) = x(s + \Delta) - x(s)$. We use $\bar{b}(x)$ to denote $\bar{H}_\alpha(x, 0, t)$, when it exists. The theorem provides a simple means of verifying Assumption 2.1(i). This is clearly consistent with our previous reference to $\bar{b}(x)$ as the "mean dynamics."

Proof. We make use of the following facts regarding convex functions. The proof of (i) is straightforward, while (ii) is Theorem 24.5 of [32].

(i) Let $\{f_i(\cdot)\}$ be convex on R^r and satisfy $f_i(0) = 0$. Let $\partial f(\alpha)$ denote the set of subdifferentials of a given convex function $f(\cdot)$ at α. Then for $f(\cdot)$ defined by $f(\alpha) = \sup_i f_i(\alpha)$, $\partial f(0)$ is the closed convex hull of $\cup_i \partial f_i(0)$.

(ii) Let $\{f_i(\cdot)\}$ be a sequence of convex functions and suppose that $f_i(\alpha) \to f(\alpha)$ in a neighborhood of $\alpha = 0$. Let $N_\varepsilon(y)$ denote the open ball of radius ε around y. Then given $\varepsilon > 0$ there is i_ε such that for $i \geqq i_\varepsilon$, $\partial f_i(0) \subset N_\varepsilon(\partial f(0))$.

We assume existence of the derivative of $\bar{H}(x, \alpha, t)$ in α at $\alpha = 0$.

Fix $\varepsilon > 0$. By using the two facts above and the definition of $\overline{\lim}$, we can obtain $\Delta_0 > 0$, $N_0 < \infty$, and $\delta > 0$ such that for all $0 < \Delta < \Delta_0$, $N \geqq N_0$, $y \in N_\delta(x)$, and $s \in N_\delta(t)$,

$$\frac{\partial}{\partial \alpha}(a_N \log E[\exp \langle \alpha, D^\Delta X^N(s) \rangle / a_N \mid \mathcal{F}_N, X^N(s) = y]/\Delta)|_{\alpha = 0} \subset N_\varepsilon(H_\alpha(x, 0, t)) \quad \text{(a.s.)}.$$

Using properties of conditional expectation to compute the derivative we obtain (for all such Δ, etc.)

$$E[D^\Delta X^N(s) \mid \mathcal{F}_N, X^N(s) = y]/\Delta \in N_\varepsilon(\bar{H}_\alpha(x, 0, t)) \quad \text{(a.s.)}.$$

By letting $\Delta \to 0$, $N \to \infty$, and $\delta \to 0$ through a subsequence we obtain (5.2). Owing to the way we define $X^N(\cdot)$ as a shifted version of $X(\cdot)$, the right-hand side of (5.2) must be independent of t (almost surely). Hence so is $\bar{H}_\alpha(x, 0, t)$. Finally, we note

that since $\bar{b}(x)$ is the unique subdifferential of $\bar{H}(x, \cdot, t)$ at $\alpha = 0$, $\bar{H}(x, \alpha, t) - \langle \alpha, \beta \rangle \geqq 0$ for all α if and only if $\beta = \bar{b}(x)$. Consequently, $\bar{L}(x, \beta, t) = 0$ if and only if $\beta = \bar{b}(x)$. □

We consider the following assumption on the sequence $\{a_n\}$.

ASSUMPTION 5.1.

$$\lim_{\substack{|t_n - t_m| \to 0 \\ n,m \to \infty}} \frac{a_n}{a_m} = 1.$$

Define $K_N(t) = a_{m(t_N+t)}/a_N$, and $K(t) = \overline{\lim}_{N\to\infty} K_N(t)$. For given $\delta > 0$ there are $c(\delta) > 0$ and $N(\delta) < \infty$ such that $N \geqq N(\delta)$ and $|t - s| \leqq c(\delta)$ imply $|K_N(t) - K_N(s)| \leqq \delta$. Thus $K(\cdot)$ is continuous, with $0 < K(t) < \infty$ for $0 \leqq t < \infty$.

Examples. Let $a_n = 1/n$. Then $m(t_n + s)/n(\exp s) \to 1$ as $n \to \infty$, and $K_N(s) \to \exp -s$. Let $a_n = 1/n^\gamma$, $\gamma \in (0, 1)$. Then $m(t_n + s)/(n + sn^\gamma) \to 1$ as $n \to \infty$ and $K_N(s) \to 1$. If $a_n = c/\log n$, then $m(t_n + s)/(n + s) \to 1$ and $K_N(s) \to 1$. In general, if a_n is nonincreasing, then $K(t) \leqq 1$.

Even if not explicitly stated, Assumptions 4.1 and 5.1 are assumed for the rest of § 5.

The next theorem shows how to simplify the calculation of $\bar{H}(x, \alpha, t)$ under Assumption 5.1 by justifying the replacement of $\{a_i\}$ by an appropriate constant sequence (for the purposes of calculating $\bar{H}$).

THEOREM 5.2. *Suppose that Assumptions* 4.1 *and* 5.1 *hold. Then*

$$\bar{H}(x, \alpha, t) \leqq K^{-1}(t) \overline{\lim_{\Delta\to 0}} \overline{\lim_{N\to\infty}} \overline{\lim_{\substack{y\to x \\ s\to t}}} \operatorname*{ess\,sup}_{\omega}$$

$$\cdot \log E\left[\exp\left\langle \alpha K(t), \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} F_i \right\rangle \middle| \mathcal{F}_N, X_{m(t_N+s)} = y\right] \Big/ \tag{5.3}$$

$$(m(t_N + s + \Delta) - m(t_N + s)).$$

Proof. Let $\delta > 0$ be given. If Δ is small and N is large, then Assumptions 4.1 and 5.1 imply

$$E\left[\exp\left\langle \alpha, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} a_i F_i \right\rangle \Big/ a_N \middle| \mathcal{F}_N, X_{m(t_N+s)} = y\right]$$

$$\leqq E\left[\exp\left\langle \alpha a_{m(t_N+s)}, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} F_i \right\rangle \Big/ a_N \middle| \mathcal{F}_N, X_{m(t_N+s)} = y\right]$$

$$\cdot \exp \delta(m(t_N + s + \Delta) - m(t_N + s)).$$

Hence

$$\bar{H}(x, \alpha, t) \leqq \overline{\lim_{\Delta\to 0}} \overline{\lim_{N\to\infty}} \overline{\lim_{\substack{y\to x \\ s\to t}}} \operatorname*{ess\,sup}_{\omega} a_N \log E$$

$$\cdot \left[\exp\left\langle \alpha a_{m(t_N+s)}, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} F_i \right\rangle \Big/ a_N \middle| \mathcal{F}_N, X_{m(t_N+s)} = y\right] \Big/ \Delta.$$

By differentiating

$$\log E\left[\exp\left\langle \alpha a_{m(t_N+s)}, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} F_i \right\rangle \Big/ a_N \middle| \mathcal{F}_N, X_{m(t_N+s)} = y\right] \Big/ \Delta \tag{5.4}$$

with respect to $a_{m(t_N+s)}$, we see that it is convex and zero if $a_{m(t_N+s)}=0$. Given a convex function $H(\alpha)$ such that $H(0)=0$, the inequality $H(s\alpha')\leqq sH(\alpha')$ is valid for all $0\leqq s\leqq 1$, and all α'. The definition of $K(\cdot)$ implies the existence of $c_N(\Delta)$ such that $c_N(\Delta)\to 0$ if $N\to\infty$ and then $\Delta\to 0$, and such that $a_{m(t_N+s)}/a_N\leqq K(s)+c_N(\Delta)$. Combining these facts we obtain the upper bound

$$\frac{a_{m(t_N+s)}}{a_N(K(s)+c_N(\Delta))} \cdot \log E\left[\exp\left\langle \alpha, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} [K(s)+c_N(\Delta)]F_i\right\rangle \middle| \mathcal{F}_N, X_{m(t_N+s)}=y\right]\Big/\Delta \tag{5.4a}$$

for (5.4). Since

$$a_{m(t_N+s)}(m(t_N+s+\Delta)-m(t_N+s))/\Delta\to 1$$

and

$$K(s)+c_N(\Delta)\to K(t)$$

as $s\to t$, $N\to\infty$, and $\Delta\to 0$, we are done. □

Remarks. We have used the fact that Assumption 5.1 implies $K_N^{-1}(t)$ is bounded from above uniformly in N for N large. Suppose all that is known about $\{a_i\}$ is that for large i the sequence is nonincreasing. If we define

$$\bar{a}_i = a_i \vee (c_i/\log i), \qquad \bar{F}_i = a_iF_i/\bar{a}_i$$

where c_i tends monotonically to zero in a "regular" way (e.g., $c_i=1/\log i$), then $\{\bar{a}_i\}$ satisfies Assumption 5.1, with $K(t)\leqq 1$. If Assumption 2.3 holds for $\{a_i\}$, then it holds for $\{\bar{a}_i\}$ as well. Hence we can apply Theorems 3.1 and 3.2 to analyze the process $X_{i+1}=X_i+\bar{a}_i\bar{F}_i$. Note that the only difference is that we are now using interpolation intervals that might be larger than the original $\{a_i\}$. Application of Theorem 5.2 yields (5.3) with F_i replaced by $\bar{F}_i=a_iF_i/\bar{a}_i$. If we can choose $\{\bar{a}_i\}$ so that $\{a_i/\bar{a}_i\}$ is nonincreasing, we can exploit convexity to obtain (5.3) as written (note, however, that $\bar{H}(x,\alpha,t)$ is now a large deviation functional for the process defined in terms of $\{\bar{a}_i\}$ and $\{\bar{F}_i\}$, and *not* $\{a_i\}$ and $\{F_i\}$).

We now consider a number of concrete examples to show that Theorems 5.1 and 5.2 can be quite convenient to use.

5.1. Random vector fields.

5.1.a. The i.i.d. case. We consider a family of probability measures parametrized by $x\in\mathbb{R}^r$, denoted by μ_x. We assume the μ_x are weakly continuous in x. The boundedness assumption implies that for every x the support of μ_x is contained in $N_K(x)$. We then consider a sequence of i.i.d. random vector fields $\{b_n(x)\}$ such that $P\{b_n(x)\in A\}=\mu_x(A)$, and set $F_n=b_n(X_n)$. Define $H(x,\alpha)=\log E\exp\langle\alpha, b_n(x)\rangle$ and $\mathcal{F}_n=\sigma(X_i, i\leqq n)$. Note that the weak continuity of μ_x implies that $H(\cdot,\cdot)$ is continuous. For the constant $a_n\equiv a$ case, and under stronger conditions on the μ_x, [27] contains a development of both the upper and lower large deviations bounds. It follows from our method that the same upper bound can be obtained more easily and under weaker conditions. (See also Example 7.1.)

THEOREM 5.3. *Under Assumptions* 4.1 *and* 5.1, $\bar{H}(x,\alpha,t)$ (*defined through* (4.1)) *is bounded above by* $K^{-1}(t)H(x,K(t)\alpha)$, *and Assumption* 2.1(i) *holds with* $\bar{b}(x)=Eb_n(x)=\int\beta\mu_x(d\beta)$.

Proof. By Theorem 5.2

$$\bar{H}(x, \alpha, t) \leqq K^{-1}(t) \overline{\lim_{\Delta \to 0}} \overline{\lim_{N \to \infty}} \overline{\lim_{\substack{y \to x \\ s \to t}}} \cdot \log E\left[\exp\left\langle \alpha K(t), \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} b_n(X_n)\right\rangle \middle| X_{m(t_N+s)} = y\right] \Big/ (m(t_N+s+\Delta) - m(t_N+s)).$$

Using the fact that $|X_n - y| \leqq K\Delta$ (almost surely) if $m(t_N+s) \leqq n \leqq m(t_N+s+\Delta)$ and $X_{m(t_N+s)} = y$, and the continuity of $H(x, \alpha)$, we can use the properties of conditional expectation to compute

$$\bar{H}(x, \alpha, t) \leqq K^{-1}(t) \overline{\lim_{\Delta \to 0}} \overline{\lim_{N \to \infty}} \sum_{m(t_N+t)}^{m(t_N+t+\Delta)} \cdot [H(x, K(t)\alpha) + O_n(\Delta)]/(m(t_N+t+\Delta) - m(t_N+t)) \tag{5.5}$$

where $O_n(\Delta) \to 0$ as $\Delta \to 0$, uniformly in n. The conclusion of the theorem obviously follows from this. □

5.1.b. The non-i.i.d. case. We consider the generalization of Section 5.1.a to the case where the $\{b_i(\cdot)\}$ are independent, but not identically distributed. We assume that the limit of

$$E \sum_{N+1}^{N+n} \frac{b_i(x)}{n}$$

exists as N and n tend to infinity (for each x). Denote the limit by $\bar{b}(x)$. We also assume that the measures induced by the $b_i(x)$ are weakly continuous in x, uniformly in i. This implies that $\bar{b}(x)$ is continuous.

Define

$$H(x, \alpha) = \overline{\lim_{N,n}} \sum_{N+1}^{N+n} \log E \exp\langle \alpha, b_i(x)\rangle / n.$$

Then the assumptions above imply that $H(\cdot, \cdot)$ is continuous and that $H_\alpha(x, 0)$ exists, which by Theorem 5.1 must equal $\bar{b}(x)$. The proof of Theorem 5.3 yields the following.

THEOREM 5.4. *Under Assumptions* 4.1 *and* 5.1, $\bar{H}(x, \alpha, t)$ *is bounded above by* $K^{-1}(t)H(x, K(t)\alpha)$, *and Assumption* 2.1(i) *holds with* $\bar{b}(x)$ *as above.*

5.2. Exogenous noise. In this section we take as our model $F_n = b(X_n, \xi_n)$, where the process $\{\xi_n\}$ is *exogenous* and takes values in a set M, i.e., for any n and Borel set $A \subset M$,

$$P\{\xi_n \in A \mid (X_i, \xi_{i-1}), i \leqq n\} = P\{\xi_n \in A \mid \xi_{i-1}, i \leqq n\} \quad \text{(a.s.)}.$$

In this section, we always assume that $b(\cdot, \xi)$ is continuous, uniformly in $\xi \in M$.

5.2.a. Markov noise. Suppose that $\{\xi_n\}$ is a Markov chain, with state space M, and with one-step transition function $P(\xi, \cdot)$. Under a uniform (in the initial condition) recurrence condition on the process $\{\xi_n\}$, the following facts are proved in [24]. Let $C(M)$ denote the continuous real-valued functions on M and define an operator mapping $C(M) \to C(M)$ by

$$\hat{P}(x, \alpha)(f)(\xi) = \int_M \exp\langle \alpha, b(x, \psi)\rangle f(\psi) P(\xi, d\psi). \tag{5.6}$$

The eigenvalue $\lambda(x, \alpha)$ of $\hat{P}(x, \alpha)$ with the maximum modulus is real, simple, and larger than unity for $\alpha \neq 0$. If we define $H(x, \alpha) = \log \lambda(x, \alpha)$, then $H(x, \alpha)$ is analytic in α, and

$$H(x, \alpha) = \lim_n \frac{1}{n} \log E_\xi \exp \left\langle \alpha, \sum_1^n b(x, \xi_i) \right\rangle, \tag{5.7}$$

uniformly in $\xi = \xi_0$. In this case, $\bar{b}(x) = \int b(x, \xi)\mu(d\xi)$, where $\mu(\cdot)$ is the unique invariant measure of $\{\xi_n\}$.

The conclusion of Theorem 5.4 holds in this case as well, with a similar proof that we now outline. Using the continuity properties of $b(\cdot, \xi)$ and the bound on $|X_n - y|$ for $m(t_N + s) \leqq n \leqq m(t_N + s + \Delta)$, and the fact that the convergence in (5.7) is uniform in ξ_0, we obtain (5.5). The conclusion follows from this.

5.2.b. Stationary *m*-dependent noise. We consider exogenous noise where $\{\xi_i\}$ is stationary and m-dependent, i.e., for any n, $\{\xi_i, i \leqq n\}$ and $\{\xi_i, i > n + m\}$ are mutually independent. We set (for $\mathcal{F}_n = \sigma(\xi_i, i \leqq n)$)

$$H(x, \alpha) = \overline{\lim_n} \operatorname*{ess\,sup}_\omega \frac{1}{n} \log E_{\mathcal{F}_0} \exp \left\langle \alpha, \sum_1^n b(x, \xi_i) \right\rangle.$$

We first prove that $H(x, \alpha)$ is "smooth" in α at $\alpha = 0$. Define $D_q(x, \alpha) = E \exp \sum_0^{q-1} \langle \alpha, b(x, \xi_i) \rangle$ and $H_q(x, \alpha) = (\log D_q(x, \alpha))/(q + m)$. By the m-dependent property and the stationarity,

$$\begin{aligned} &\frac{1}{kq + km} \log E_{\mathcal{F}_{-m}} \exp \sum_1^{kq+km} \langle \alpha, b(x, \xi_i) \rangle \\ &\quad \leqq \frac{1}{kq + km} \log \left([\exp |\alpha| Kkm] \sum_{l=1}^{k} E_{\mathcal{F}_{lq-m}} \exp \sum_{lq}^{lq+q-1} \langle \alpha, b(x, \xi_i) \rangle \right) \\ &\quad \equiv \delta_q |\alpha| + H_q(x, \alpha) \end{aligned}$$

where $\delta_q = Km/(q + m) \to 0$ as $q \to \infty$. Thus, $H(x, \alpha) \leqq H_q(x, \alpha) + \delta_q |\alpha|$. At $\alpha = 0$, the gradient of $H_q(x, \alpha)$ equals $E \sum_1^q b_i(x, \xi_i)/(q + m)$, which converges to a limit $\bar{b}(x)$ as $q \to \infty$. Since the convex (in α) function $H(x, \alpha)$ is bounded above by the convex functions $H_q(x, \alpha) + \delta_q |\alpha|$, and since $H(x, 0) = H_q(x, 0) = 0$, the set of subdifferentials of $H(x, \cdot)$ at $\alpha = 0$ is contained in the set of subdifferentials of $H_q(x, \cdot) + \delta_q |\cdot|$ for every q. This latter set converges to the point $\bar{b}(x)$ as $q \to \infty$. Hence $H(x, \cdot)$ has $\bar{b}(x)$ as its unique subdifferential at $\alpha = 0$, which implies that $H_\alpha(x, 0)$ exists and equals $\bar{b}(x)$.

Using these facts and the same argument as that in § 5.2.a, we again obtain the conclusion of Theorem 5.4.

5.3. State-dependent noise. To model state-dependency effects, we consider a model of the form $F_n = b_n(X_n, \xi_n)$, such that the pair (X_n, ξ_{n-1}) is Markovian. We assume the following:

(1) $\{b_i(\cdot, \cdot)\}$ is a sequence of i.i.d. random vector fields;

(2) There is a transition kernel $P^x(\xi, \cdot)$ such that

$$P\{\xi_n \in \cdot \mid \xi_{n-1} = \xi, X_n = x, \xi_{i-1}, X_i, i < n\} = P^x(\xi, \cdot).$$

This approach to modeling state-dependent effects includes some of the models of [28], [29], and [31] that restrict $\{b_i\}$ to the case of a deterministic vector field. An extensive study of lower and upper large deviations bounds for a more restricted version of such models is in [20].

In addition to the "true" noise process, we will make use of a fixed-x Markov process $\{\xi_i^x\}$ that is simply the process generated by the kernel $P^x(\cdot,\cdot)$.

Let $P^{x,n}(\xi, d\psi)$ denote the n-step transition kernel of $\{\xi_i^x\}$. Then we have the following basic assumptions on the model.

(i) There exist $n_0 < \infty$ and $\delta > 0$ (which may depend on x) such that for all Borel $A \subset M$,

$$\inf_{\xi \in M} P^{x,n_0}(\xi, A) \geqq \delta \sup_{\xi \in M} P^{x,n_0}(\xi, A).$$

(ii) $E \int_M \exp\langle \alpha, b_n(x, \psi)\rangle P^x(\xi, d\psi)$ is continuous in x, uniformly in ξ.

Remarks. Assumption (i) implies that there is a uniform (in ξ) rate at which the measures induced by ξ_i^x converge to the invariant measure, given $\xi_0^x = \xi$. There are obviously many sets of sufficient conditions that yield (ii). It is also possible (though notationally cumbersome) to consider a weaker form of (ii) that requires the existence of $n < \infty$ such that

$$E \int\int \cdots \int \exp\langle \alpha, b_n(x, \psi)\rangle P^{x_0}(\xi, d\xi_1) P^{x_1}(\xi_1, d\xi_2) \cdots P^{x_n}(\xi_n, d\psi)$$

is continuous in $(x_0, \cdots, x_n)$, uniformly in $\xi \in M$.

The basic method for proving that an analogue of Theorem 5.4 holds for this model is to first show that something like (5.7) holds for the fixed-x process, and to then use the continuity properties assumed in (ii) above. The analogue of (5.6) for this case will be

$$\hat{P}(x, \alpha)(f)(\xi) = E \int_M \exp\langle \alpha, b_i(x, \psi)\rangle f(\psi) P^x(\xi, d\psi).$$

Under the above assumptions, the results in [24] apply here as well, and we find that the limit

$$H(x, \alpha) = \lim_n \frac{1}{n} \log E_\xi \exp\langle \alpha, b_i(x, \xi_i^x)\rangle \tag{5.8}$$

exists uniformly in $\xi = \xi_0^x$, where $H(x, \alpha)$ is continuous in x and analytic in α. The assumptions also imply that for each x, the process $\{\xi_i^x\}$ has a unique invariant measure $\mu^x(\cdot)$, and that

$$\bar{b}(x) = H_\alpha(x, 0) = E \int b_i(x, \xi) \mu^x(d\xi).$$

Proof of Theorem 5.4 (*for the present case*). By (i) and (ii) above, (5.8) holds [24], where once again (see § 5.2) $H(x, \alpha)$ is defined as the log of an eigenvalue of the operator $\hat{P}(x, \alpha)$. To simplify the notation, we set $n_0 = 1$. The general proof is very similar. Let F be a fixed compact set. Using Theorem 5.2 (as was done in Theorem 5.3), it is sufficient to restrict our attention to estimates on terms of the form (5.9) below, where $|X_i - y| \leqq \delta$ for all $1 \leqq i \leqq n$. The transition kernel used to get the ξ_n in $b_n(X_n, \xi_n)$ will be $P^{X_n}(\xi_{n-1}, \cdot)$. Then for $\xi = \xi_0$,

$$\begin{aligned} & E_\xi \exp\left\langle \alpha, \sum_1^n b_i(X_i, \xi_i)\right\rangle \\ & = E_\xi\left[\exp\left\langle \alpha, \sum_{i=1}^{n-1} b_i(X_i, \xi_i)\right\rangle E_\xi[\exp\langle \alpha, b_n(X_n, \xi_n)\rangle \mid \xi_{n-1}, X_n]\right] \\ & = E_\xi\left[\left[\exp\left\langle \alpha, \sum_{i=1}^{n-1} b_i(X_i, \xi_i)\right\rangle\right]\right. \\ & \quad \left. \cdot E_\xi\left[\int_M \exp\langle \alpha, b_n(X_n, \psi)\rangle P^{X_n}(\xi_{n-1}, d\psi) \mid \xi_{n-1}, X_n\right]\right]. \end{aligned} \tag{5.9}$$

Under (ii) above, there is $\gamma(\delta) > 0$ such that $\gamma(\delta) \to 0$ as $\delta \to 0$, and such that the last bracketed terms are bounded above by

$$E_\xi \int_M \exp \langle \alpha, b_n(y, \psi) \rangle P^y(\xi_{n-1}, d\psi) \exp \gamma(\delta). \tag{5.10}$$

Using (5.10) in (5.9) and continuing to iterate backward to approximate all the X_i by y plus an "error," we obtain the upper bound to (5.9) of

$$E_\xi \exp \left[\left\langle \alpha, \sum_1^n b_i(y, \xi_i^y) \right\rangle + n\gamma(\delta) \right].$$

The proof is now essentially the same as that of Theorem 5.3 save that we use (5.8). The σ-algebra here should be $\mathcal{F}_n = \sigma(X_i, \xi_{i-1}, i \leqq n)$.

5.4. A level-2 and level-3 large deviation approach. We have seen that when the state process $\{X_n\}$ is Markov, or when it is one component of a Markov process (such as our "state-dependent noise" process $\{X_n, \xi_{n-1}\}$) satisfying certain "uniformly recurrent" conditions, it is possible to prove the differentiability of $\bar{H}(x, \alpha, t)$ for a wide class of such processes by using analytical techniques and the characterization of the time-independent (constant gain) version $H(x, \alpha)$ as the log of the eigenvalue of largest modulus of an operator associated to the process (as in (5.6)). We again refer the reader to [24] for details on how this approach may be used in a general setting. However, for many of the processes arising in the study of stochastic systems, the assumptions required by this approach do not hold. As a very common example, we may consider the ARMA model to be discussed below.

In this section, we outline a method for proving the differentiability of $H(x, \alpha)$ at $\alpha = 0$ (which, as we have seen, will imply that of $\bar{H}(x, \alpha, t)$) that is based on well-known "level 2" and "level 3" large deviations results and that is general enough to cover many of the non-Markov processes encountered in recursive algorithms.

In applications, we would not want to be concerned with the abstract level of results in this section. But, they make it clear that the α-differentiability assumption is not restrictive and can be treated in many different ways. We work only with the exogenous noise case and $F_n = b(X_n, \xi_n)$ and with $b(\cdot, \cdot)$ bounded for simplicity of exposition. Define the sample occupation measure (over the Borel sets Γ) by

$$L_N(\Gamma, \omega) = \frac{1}{N} \sum_1^N I_{\{b(x, \xi_i) \in \Gamma\}},$$

and the space $\mathcal{M}$ as the set of probability measures on R^r, endowed with the topology of weak convergence. The $L_N(\cdot, \omega)$ are in $\mathcal{M}$. Assume the following large deviations estimate (x is fixed throughout):

ASSUMPTION 5.2. There is a lower-semicontinuous nonnegative functional I_x on $\mathcal{M}$ such that the sets $\{\gamma \in \mathcal{M}: I_x(\gamma) \leqq s\}$ are compact for $s < \infty$, and for Borel $A \subset \mathcal{M}$ and each ω (w.p.1)

$$\overline{\lim_N} \frac{1}{N} \log P_{\mathcal{F}_0}\{L_N \in A\} \leqq -\inf_{v \in \bar{A}} I_x(v). \tag{5.11}$$

Sufficient conditions for this assumption and Assumption 5.3 below are contained in many places (e.g., [22], [23]).

ASSUMPTION 5.3. There is a unique measure $\bar{v}_x \in \mathcal{M}$ such that $I_x(\bar{v}_x) = 0$.

It follows that $L_N(\cdot, \omega)$ converges (w.p.1) to $\bar{v}_x(\cdot)$.

Remark. Instead of Assumption 5.2 we can consider a weaker version (that we call Assumption 5.2^e). Under this assumption we require that there exist $\Omega_N \subset \Omega$ such that $\overline{\lim}_N a_N \log P(\Omega \backslash \Omega_N) = -\infty$ and such that (5.11) holds uniformly for $\omega \in \Omega_N$. Under Assumptions 5.2^e and 5.3, the arguments below and Theorem 4.1 yield Assumption 2.1^e (see the remark following the proof of Theorem 4.1).

We now show that Assumptions 5.2 and 5.3 imply the desired α-differentiability. Then an example will be given, and the approach discussed.

By Varadhan's Theorem on the asymptotic evaluation of integrals [21] and the boundedness and continuity of $b(x, \cdot)$, the following inequality holds (w.p.1):

$$\begin{aligned} \overline{\lim_N} \frac{1}{N} \log E_{\mathcal{F}_0} \exp \left\langle \alpha, \sum_1^N b(x, \xi_i) \right\rangle &\leqq \sup_v \left[\int \langle \alpha, y \rangle v(dy) - I_x(v) \right] \\ &\equiv H^*(x, \alpha). \end{aligned} \tag{5.12}$$

Define $H(x, \alpha)$ as the left-hand side of (5.12). Obviously, $H(x, \alpha) \leqq H^*(x, \alpha)$. Since both functions are convex and $H(x, 0) = H^*(x, 0) = 0$, $H(x, \alpha)$ is α-differentiable at $\alpha = 0$ if $H^*(x, \alpha)$ is.

Next note that

$$\begin{aligned} H^*(x, \alpha) &= \sup_\beta \sup_{\{v \in M : \int y v(dy) = \beta\}} \left[\int \langle \alpha, y \rangle v(dy) - I_x(v) \right] \\ &= \sup_\beta \left[\langle \alpha, \beta \rangle - L^*(x, \beta) \right] \end{aligned} \tag{5.13}$$

where

$$L^*(x, \beta) = \inf_{\{v \in \mathcal{M} : \int y v(dy) = \beta\}} I_x(v).$$

Since $H^*(x, 0) = 0$, β^* is a subdifferential of $H^*(x, \alpha)$ at $\alpha = 0$ if and only if

$$H^*(x, \alpha) - \langle \alpha, \beta^* \rangle \geqq 0 \quad \text{for all } \alpha. \tag{5.14}$$

But (5.14) holds if and only if $L^*(x, \beta^*) = 0$ since H^* is the Legendre transform of L^*. Since $H^*(x, \cdot)$ is convex, it is differentiable at $\alpha = 0$ if and only if the set of subdifferentials at $\alpha = 0$ contains only one element. By Assumption 5.3, $\beta^* = \int y \bar{v}_x(dy)$ is the unique value of β for which $L^*(x, \beta) = 0$. Thus $\beta^* = \int y \bar{v}_x(dy)$ is the unique subdifferential, and the α-differentiability is proved. Note that $\beta^* = \bar{b}(x)$, as defined in Theorem 5.1.

Discussion. We have phrased our requirement in terms of the differentiability of $H(x, \alpha)$ at $\alpha = 0$, but as shown above this is obviously *equivalent to the uniqueness of the β^* satisfying $L(x, \beta^*) = 0$*. The reason for our choice is that in most of the work on large deviations for dynamical systems [12], as well as the work generalizing Cramer's original paper [16], [24], [25], the differentiability of $H(x, \alpha)$ in α is taken as a fundamental assumption. As a consequence, this was the condition that was typically verified for a given noise process.

We illustrate the method with an example.

Example. Suppose that $b(x, \cdot)$ is continuous, and that $\{\xi_n\}$ is a stationary ARMA process with representation

$$A_0 \xi_n + A_1 \xi_{n-1} + \cdots + A_{d_1} \xi_{n-d_1} = B_0 \psi_n + B_1 \psi_{n-1} + \cdots + B_{d_2} \psi_{n-d_2}$$

where $\{\psi_i\}$ is a sequence of zero mean, bounded, i.i.d. random variables. For simplicity, we assume both ξ_i and ψ_i take values in R^r. It is also assumed that the roots of $\det(A_0 + A_1 s + \cdots + A_{d_1} s^{d_1})$ lie outside of the closed unit disc.

Define $S = (R^r)^z$ (the space of bounded infinite sequences $\{q_i, -\infty < i < \infty\}$ with values in R^r), and consider the mapping $F: S \to S$ defined by $(F(\ )_j$ denotes the jth component)

$$F(\{s_i\})_j = b(x, p_j)$$

where $\{s_i\}$ and $\{p_i\}$ are related by

$$A_0 p_n + \cdots + A_{d_1} p_{n-d_1} = B_0 s_n + \cdots + B_{d_2} s_{n-d_2}, \qquad -\infty < n < \infty.$$

This relation defines $\{p_n\}$ uniquely in terms of $\{s_n\}$. We can metrize S in such a way that F is continuous (and in fact uniformly continuous on a subset $A \subset S$ such that $\{\psi_i\} \subset A$ (w.p.1)). (For example, for $0 < \lambda < 1$, use the metric $d(\{s_j\}, \{s'_j\}) = \sum_{-\infty}^{\infty} \lambda^{|j|} \min\{1, |s_j - s'_j|\}$.) It is then relatively straightforward to show that Assumptions 5.2 and 5.3 follow from the (so-called "level 3") large deviations results for the process $\{\psi_i\}$ that are given in [23], under a suitable application of the "contraction principle" (a "continuous mapping" technique) [21, § 2]. We omit all details here, since they would take us too far afield, and the techniques are known in large deviations theory.

In general, if a given process $\{\xi_i\}$ can be represented as a continuous transformation of a simpler process $\{\psi_i\}$ for which the appropriate "level 3" results exist, then we may obtain Assumptions 5.2 and 5.3 via the "contraction principle."

Although this approach may seem abstract, it, in fact, easily yields the α-differentiability for a wide variety of the noise processes of interest in stochastic systems theory, which often *do* have such a representation.

6. The constrained algorithm. In this section we show that Assumption 2.1 holds for the "constrained" (or "projected") version of any of the models considered in § 5.

Let G be a bounded open set contained in $\mathbb{R}^r$. We define the set of exterior normals at $x \in \partial G$ by

$$n(x) = \{n' \in \mathbb{R}^r: |n'| = 1, \exists c > 0 \text{ such that } \langle x - y, n' \rangle + (1/2c)|x - y|^2 \geqq 0 \text{ for } y \in G\}.$$

It follows that $n' \in n(x)$ if and only if there is $c > 0$ such that $N_c(x + cn') \cap G = \varnothing$. We make the following assumptions on the set G.

ASSUMPTION 6.1. (1) The boundary of G is the union of a countable number of smooth (C^2) surfaces.

(2) G satisfies a "uniform exterior sphere condition": there is $c_0 > 0$ such that for all $x \in \partial G$ and all $n' \in n(x)$, $N_{c_0}(x + c_0 n') \cap G = \varnothing$.

(3) There are $\delta > 0$ and $\tau \in (0, 1]$ such that given any $x \in \partial G$ there is a unit vector $l(x)$ such that $\langle l(x), n'' \rangle \geqq \tau$ for all $n'' \in \{n' \in n(y): y \in N_\delta(x) \cap \partial G\}$.

Remarks. Part (2) is satisfied if G is convex, or if G has a piecewise smooth boundary with "convex corners." Part (3) is *Condition* (B) of [17], and is implied by a "*uniform interior cone condition*" (see [17]). In particular, (3) is satisfied if G is convex and bounded, with nonempty interior.

It is well known, under these conditions on G, that to each point x in the c_0-neighborhood of G we may associate a unique closest point $\pi_G(x)$ in $\bar{G}$. Furthermore, if $x \notin G$, then there is $\gamma \geqq 0$ and $n' \in n(\pi_G(x))$ such that $x - \pi_G(x) = \gamma n'$.

All the models of § 5 fall into either of two classes. For those processes in the first class we have $F_n = b_n(X_n, \xi_n)$, where $\{b_n(\cdot, \cdot)\}$ is a sequence of independent vector

fields, and the process $\{\xi_n\}$ is *exogenous.* Recall that by "exogenous," we mean that for all Borel A and integer n,

$$P\{\xi_n \in A \mid X_i, \xi_{i-1}, i \leqq n\} = P\{\xi_n \in A \mid \xi_{i-1}, i \leqq n\}.$$

In this case the "projected" version of (2.1) is defined by

$$X_{n+1} = \pi_G(X_n + a_n b_n(X_n, \xi_n)). \tag{6.1}$$

Alternatively, we have considered models such that the $\{b_n(\cdot, \cdot)\}$ is as above but where the pair (X_n, ξ_{n-1}) is Markov, with the distribution of ξ_n given $X_n = x$ and $\xi_{n-1} = \xi$ specified by the kernel $P^x(\xi, \cdot)$. We define the constrained version of this type of model by defining X_{n+1} via (6.1), and then generating ξ_{n+1} according to the kernal $P^{X_{n+1}}(\xi_n, \cdot)$.

With the "projected" version defined in this way, we use $X(\cdot)$ and $X^N(\cdot)$ to denote the usual piecewise linearly interpolated and shifted versions. For each N we also define an "unprojected" version:

$$\tilde{X}_n^N = \sum_N^{n-1} a_i b_i(X_i, \xi_i) + X_N, \qquad n \geqq N. \tag{6.2}$$

Note that $\{X_i, i \geqq N\}$ can be constructed from $\{\tilde{X}_i^N, i \geqq N\}$. We let $\tilde{X}^N(\cdot)$ denote the piecewise linear interpolated version of this process that starts at X_N at time $t = 0$, and uses interpolation intervals $\{a_i, i \geqq N\}$.

The method of proving that the appropriate analogue of Assumption 2.1 holds for $X^N(\cdot)$ involves first proving a large deviation upper bound for $\tilde{X}^N(\cdot)$, and then "transferring" it to $X^N(\cdot)$. Large deviations upper bounds are obtainable from §§ 4 and 5, since $|X_{i+1} - X_i| = O(a_i) = |\tilde{X}_{i+1}^N - \tilde{X}_i^N|$. To "transfer" such a bound to the constrained algorithm, an appropriate "continuous mapping" between $\tilde{X}^N(\cdot)$ and $X^N(\cdot)$ is needed. This requires some additional definitions, but they are only intermediaries in the proof.

We next define a collection of mappings $\{C_i^N\}$ on sequences $\{x_j\}$, $x_j \in \mathbb{R}^r$. For each N, and $i \geqq N$, we define the function $C_i^N(\cdot)$ and the vectors $\bar{x}_i$ recursively by $\bar{x}_N = x_N = C_N^N(\{x_j\})$, and for $i \geqq N$, $\bar{x}_{i+1} = C_{i+1}^N(\{x_j\}) = \pi_G(\bar{x}_i + (x_{i+1} - x_i))$. With this definition, the sequence $\{\tilde{X}_i\}$ defined by (6.2) can also be written as

$$\tilde{X}_{i+1}^N = \tilde{X}_i^N + a_i b_i(C_i^N(\{\tilde{X}_j^N, j = N, \cdots, i\}), \xi_i).$$

The maps C_i^N are obviously continuous in the following sense. Given any sequence $\{x_i, i \geqq N\}$ such that $|x_{i+1} - x_i| \leqq c_0/2$ for all $i \geqq N$, we have (for $i_2 \geqq i_1$)

$$|C_{i_1}^N(\{x_j\}) - C_{i_2}^N(\{x_j\})| \leqq B \sum_{i_1}^{i_2 - 1} |x_{j+1} - x_j| \tag{6.3}$$

where B is the Lipschitz constant of $\pi_G(\cdot)$ on the $c_0/2$-neighborhood of G.

The appropriate definition of $\bar{H}(x, \alpha, t)$ for this case is

$$\bar{H}(x, \alpha, t) = \overline{\lim_{\Delta \to 0}} \overline{\lim_{N \to \infty}} \overline{\lim_{\substack{y \to x \\ s \to t}}} \operatorname*{ess\,sup}_{\omega} a_N \cdot \log E[\exp \langle \alpha, D^\Delta \tilde{X}^N(s) \rangle / a_N \mid \mathcal{F}_N, X_{m(t_N + s)} = y]/\Delta.$$

Note that y in this expression is the value the "true" projected version takes at time $m(t_N + s)$.

Important Remark. For any of the models of § 5 the function $\bar{H}$ so defined is the same as that of the corresponding "unprojected" version.

This fact follows from the proofs for the "unprojected" versions, simply by making use of (6.3). In particular, since $\bar{H}_\alpha(x, 0, t)$ exists for the "unprojected" version of each of the models of § 5, it exists here as well. We define $\bar{L}(x, \beta, t)$ as usual.

Before stating the theorem we must define the appropriate $\bar{L}$ and $\bar{S}$ functionals for the constrained problem. For $x \in \bar{G}$, and $v \in \mathbb{R}^r$, we set

$$\pi_G(x, v) = \lim_{\Delta \downarrow 0} [\pi_G(x + \Delta v) - x]/\Delta. \tag{6.4}$$

Then [26, Lemma 4.6] $\pi_G(x, v)$ equals v if $x \in G^0$ (the interior of G) or if $x \in \partial G$ and $\sup_{\gamma \in n(x)} \langle \gamma, v \rangle < 0$ (i.e., where v points inward). In general, it equals $v - \langle v, \gamma^* \rangle \gamma^*$ if $x \in \partial G$, $\sup_{\gamma \in n(x)} \langle \gamma, v \rangle \geqq 0$ and γ^* is a maximizer.

We now define the "constrained" L-functional by

$$\bar{L}_G(x, \beta, t) = \inf_{v : \pi_G(x,v) = \beta} \bar{L}(x, v, t). \tag{6.5}$$

For $x \notin \bar{G}$ or if the infimizing set is empty, set $\bar{L}_G(x, \beta, t) = +\infty$. Then $\bar{L}_G(x, \beta, t) = \bar{L}(x, \beta, t)$ if $x \in G^0$ or if $x \in \partial G$ and $\langle \gamma, \beta \rangle < 0$ for all $\gamma \in n(x)$ (i.e., β points to the interior of G). If $x \in \partial G$ and there is $\gamma \in n(x)$ such that $\langle \gamma, \beta \rangle > 0$ (β points "out" of G), then $\bar{L}_G(x, \beta, t) = \infty$. The interesting case is when $\sup_{\gamma \in n(x)} \langle \gamma, \beta \rangle = 0$; i.e., β points "along the boundary." In this case, there is a true (nontrivial) minimization. Since $\bar{L}(x, \beta, t)$ is l.s.c. in β and $\bar{L}(x, \beta, t) \to \infty$ as $|\beta| \to \infty$, the infima is attained. Define

$$\bar{S}_G(x, T, \phi) = \int_0^T \bar{L}_G(\phi(s), \dot{\phi}(s), s)\, ds, \tag{6.6}$$

and the ODE for the projected mean dynamics

$$\dot{x} = \pi_G(x, \bar{b}(x)) \tag{6.7}$$

where $\bar{b}(\cdot)$ is defined by $\bar{H}_\alpha(x, 0, t)$.

One of the main difficulties as well as points of interest for the constrained algorithm is that in many applications the escape of $\{X_n\}$ from a neighborhood of a stable point of (6.7) will be essentially along the boundary, and when such neighborhoods are entered from the outside it is often essentially along the boundary as well.

THEOREM 6.1. *Consider the "projected" form* (6.1) *of any of the models of* § 5. *Then under Assumption* 6.1 *on the set G, Assumption* 2.1 *holds, but with* $\bar{b}(x)$ *replaced by* $\pi_G(x, \bar{b}(x))$, *and* $\bar{S}(x, T, \phi)$ *replaced by* $\bar{S}_G(x, T, \phi)$.

It follows that the essential large deviation assumption required by Theorems 3.1 and 3.2 is as easy to obtain for the "projected" algorithm as for the "unprojected" algorithm, under Assumption 6.1.

Proof. Let J be the set of Lipschitz continuous (constant K) paths starting in G. For $y(\cdot) \in J$, define $y_i^N = y(t_i - t_N)$ for $i \geqq N$, and $z_i^N = C_i^N(\{y_j^N, j \geqq N\})$. Let $z^N(\cdot)$ denote the usual piecewise linear interpolation of $\{z_i^N\}$, $i \geqq N$, with interpolation intervals $\{a_i, i \geqq N\}$. We make the following claim. Considered as functions on $[0, T]$, $z^N(\cdot) \to z(\cdot)$ uniformly, where

$$\dot{z} = \pi_G(z, \dot{y}), \qquad z(0) = y(0). \tag{6.8}$$

We denote this relationship by $z(\cdot) = C(y(\cdot))$. Furthermore, this convergence is uniform in $y(\cdot) \in J$, and $C(y(\cdot))$ is continuous in $y(\cdot) \in J$.

Before proving this claim, we give the proof of the theorem. When we assume the claim, the same proof as that of Theorem 4.1 (part (iv)) yields that $\tilde{X}^N(\cdot)$ satisfies a large deviation upper bound (in the sense of Assumption 2.1(iv)), with functional

$$\bar{S}^*(x, T, \phi) = \int_0^T \bar{L}(C(\phi)(s), \dot{\phi}(s), s)\, ds.$$

Using the claim again, we obtain

$$\sup_{t\in[0,T]} |X^N(t) - C(\tilde{X}^N(\cdot))(t)| \to 0$$

uniformly, w.p.1. Since $C(\cdot)$ is continuous (see the comment concerning this in the proof of the claim below), $X^N(\cdot)$ satisfies an upper bound (in the sense of Assumption 2.1(iv)) by the Proof of Theorem 3.3.1 [15], with functional

$$\bar{S}_G(x, t, \phi) = \inf_{\psi:\, \phi = C(\psi)} \bar{S}^*(x, T, \psi).$$

Due to our Assumption 6.1 on G, this $\bar{S}_G$ is exactly the same as defined by (6.5) and (6.6) [26, Lemma 4.7].

The remaining properties given by Assumption 2.1 follow easily from the definition of $\bar{S}_G$. Part (i) is obvious, as is (ii) due to the definition of $\bar{L}_G(x, \beta, s)$, while (iii) is immediate from the continuity of $C(\cdot)$ and the compactness of the level sets of $\bar{S}^*$.

Finally, we consider the proof of the claim. Let $\bar{y}^N(t) = y_i^N$ and $\bar{z}^N(t) = z_i^N$ for $t \in [t_i - t_N, t_{i+1} - t_N)$. Then the pair $(\bar{z}^N(\cdot), \bar{z}^N(\cdot) - \bar{y}^N(\cdot))$ comprise a solution of the *Skorokhod Problem* (for the path $\bar{y}^N(\cdot)$ and domain G) in the sense of [17, § 1]. It follows from Theorem 4.1 of [17] that $\bar{z}^N(\cdot) \to z(\cdot)$ (where $z(\cdot)$ is defined by (6.8)) uniformly on $[0, T]$, and that $z(\cdot)$ depends continuously on $y(\cdot)$. A review of the method of proof employed in [17] also shows that the convergence is uniform for all $y(\cdot)$ satisfying a common Lipschitz condition and starting in G. This completes the proof. □

7. Unbounded noise. An analogue of Theorem 4.1 holds for unbounded noise as well, under an assumption on the "tails" of the noise.

ASSUMPTION 7.1. There exist a σ-algebra $\mathcal{F}_n \supset \sigma(X_i, i \leqq n)$, $\gamma > 1$, and $B < \infty$ such that for all n and $s \geqq 0$,

$$P\{|F_n| \geqq s \,|\, \mathcal{F}_n\} \leqq B \exp -s^\gamma \quad \text{a.s.} \tag{7.1}$$

THEOREM 7.1. *Under Assumptions* 2.3, 5.1, *and* 7.1, *the conclusion of Theorem* 4.1 *holds.*

Proof. Assumption 2.1(ii) again follows from $\bar{H}(x, 0, t) = 0$. Under Assumption 7.1, there is a fixed convex function in $H(\alpha)$ that takes a finite value for each $\alpha \in \mathbb{R}^r$, and such that $\bar{H}(x, \alpha, t) \leqq H(\alpha)$ for all (x, t). By convex duality, $\bar{L}(x, \beta, t) \geqq L(\beta)$, where L is the Legendre transform of H. Since $H(\alpha)$ is finite for each α, $L(\cdot)$ grows faster than linearly: $L(sv)/s \to +\infty$ as $s \to +\infty$, for all $v \in \mathbb{R}^r$ with $|v| \neq 0$. The lower-semicontinuity of $\bar{L}$ and this fact are sufficient for Assumption 2.1(iii) [14].

Finally we consider part (iv). If it is demonstrated that, for each $M_1 < \infty$, $T < \infty$ and compact F, there is a compact set of paths J in $C[0, T]$ such that for all $x \in F$

$$P\{X^N(\cdot) \notin J \,|\, X_N = x, \mathcal{F}_N\} \leqq \exp -M_1/a_N \quad \text{(a.s.)} \tag{7.2}$$

when N is sufficiently large, then the same proof as that of Theorem 4.1 applies, since we can effectively ignore sample paths outside of J.

To prove (7.2), we let a compact set $F \subset \mathbb{R}^r$ be given and define $J(\delta)$, $\delta > 0$, as the compact set of paths $\{\phi(\cdot)\}$ that start in F at time zero and satisfy

$$\sup_{\substack{|s-t| \leqq \varepsilon \\ s,t \in [0,T]}} |\phi(s) - \phi(t)| \leqq \varepsilon^\delta. \tag{7.3}$$

It follows that for $x \in F$ and w.p.1,

$$
\begin{aligned}
&P\{X^N(\cdot) \notin J(\delta) \mid X_N = x, \mathcal{F}_N\} \\
&\quad \leqq P\{|F_i| \geqq a_i^{\delta-1}, \text{ some } i \in [N, m(t_N + T)] \mid X_N = x, \mathcal{F}_N\} \\
&\quad \leqq \sum_{N}^{m(T+t_N)} B \exp -a_i^{\gamma(\delta-1)} \\
&\quad \leqq B \exp -M_1/a_N \left[\sum_{N}^{m(T+t_N)} \exp\left(-a_i^{\gamma(\delta-1)} + M_1 a_N^{-1}\right) \right].
\end{aligned} \tag{7.4}
$$

We are finished if we prove that the bracketed term at the end of (7.4) is bounded for large N. Choose $\delta > 0$ such that $\gamma(\delta - 1) < -1$ (this is possible since $\gamma > 1$). By Assumption 5.1, $a_i^{\gamma(\delta-1)} \geqq 2M_1 a_N^{-1}$ for large N and for all $i \geqq N$. Hence the bracketed term is obviously bounded (for large N) by Assumption 2.3. □

Remarks. It is not actually necessary to use Assumption 2.3 to bound the bracketed term, since we sum over a finite number of indices. This is important in (for example) the constant gain analogues of the processes considered here (such as in [12]) where a_i is replaced by a fixed value $\varepsilon > 0$ and we consider the behavior of the linearly interpolated process (with interpolation interval ε) as $\varepsilon \to 0$. Here $m(T + t_N) - N = T/\varepsilon$, and we are then interested in bounding

$$(T/\varepsilon) \exp\left(-\varepsilon^{\gamma(\delta-1)} + M_1 \varepsilon^{-1}\right).$$

This quantity is obviously bounded as $\varepsilon \to 0$ if we choose δ so that $\gamma(\delta - 1) < -1$. If instead of requiring (7.1) to hold almost surely we only require that it hold for all $\omega \in \Omega_N$ and $N \leqq n \leqq m(t_N + T)$, where Ω_N is such that $\overline{\lim}\, a_N \log P(\Omega \backslash \Omega_N) = -\infty$, then we obtain the conclusion of Theorem 4.1 with Assumption 2.1^e replacing Assumption 2.1.

Theorem 5.1 does not require boundedness of $\{F_i\}$, and so holds here as well. The proof of Theorem 5.2 must be modified slightly.

THEOREM 7.2. *Suppose Assumptions* 5.1 *and* 7.1 *hold. Then the conclusion of Theorem* 5.2 *follows.*

Proof. For $\delta > 0$ define

$$F_i^\delta = \begin{cases} F_i, & |F_i| \leqq 1/\delta, \\ F_i/|F_i|\delta, & |F_i| > 1/\delta. \end{cases}$$

Fix $\alpha \in \mathbb{R}^r$. Using Assumption 7.1 and the bound (under Assumption 5.1) on a_i/a_N for $N \leqq i \leqq m(t_N + T)$, there is $c(\delta) \to 0$ as $\delta \to 0$ such that for all such i

$$E[\exp\langle \alpha, a_i F_i\rangle / a_N \mid \mathcal{F}_i] \leqq E[\exp\langle \alpha, a_i F_i^\delta\rangle / a_N \mid \mathcal{F}_i] \exp c(\delta) \tag{7.5}$$

and

$$E[\exp\langle \alpha K(t), F_i^\delta\rangle \mid \mathcal{F}_i] \leqq E[\exp\langle \alpha K(t), F_i\rangle \mid \mathcal{F}_i] \exp c(\delta) \tag{7.6}$$

almost surely.

Using properties of conditional expectation, we have

$$
\begin{aligned}
\bar{H}(x, \alpha, t) \leqq \overline{\lim_{\Delta \to 0}}\, \overline{\lim_{N \to \infty}}\, \overline{\lim_{\substack{y \to x \\ s \to t}}}\, \operatorname*{ess\,sup}_{\omega} a_N \\
\cdot \log E\left[\exp\left\langle \alpha, \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} a_i F_i^\delta \right\rangle \Big/ a_N \mid \mathcal{F}_N, X^N(s) = y \right] \Big/ \Delta + c'(\delta)
\end{aligned}
$$

where

$$c'(\delta) = \overline{\lim_{\Delta\to 0}}\ \overline{\lim_{N\to\infty}}\ \overline{\lim_{s\to t}}\ a_N(m(t_N+s+\Delta)-m(t_N+s))c(\delta)/\Delta$$

tends to zero as $\delta \to 0$. Applying Theorem 5.2, we obtain

$$\begin{aligned}\bar{H}(x,\alpha,t) \leqq K^{-1}(t)\ &\overline{\lim_{\Delta\to 0}}\ \overline{\lim_{N\to\infty}}\ \overline{\lim_{\substack{y\to x\\ s\to t}}}\ \operatorname*{ess\,sup}_{\omega} \log E \\ &\cdot\left[\exp\left\langle \alpha K(t), \sum_{m(t_N+s)}^{m(t_N+s+\Delta)} F_i^{\delta}\right\rangle \,\middle|\, \mathcal{F}_N, X_{m(t_N+s)} = y\right] \\ &\div (m(t_N+s+\Delta)-m(t_N+s)) + c'(\delta).\end{aligned} \tag{7.7}$$

Using (7.6), we obtain (7.7) with F_i^{δ} replaced by F_i, and $c'(\delta)$ replaced by $c'(\delta)+c''(\delta)$, $c''(\delta) = K^{-1}(t)c(\delta)$. Since $\delta > 0$ is arbitrary, we obtain (5.3). □

Example 7.1. We consider the extension of the model of § 5.1.a to the case where the support of μ_x is possibly unbounded. We assume that μ_x is weakly continuous in x, and that there are $\gamma > 0$ and $B < \infty$ such that

$$\int I_{\{y:|y|\geqq s\}}\,\mu_x(dy) \leqq B\exp -s^{\gamma} \tag{7.8}$$

for all x.

From Theorem 7.1 we obtain (under Assumption 5.1) parts (ii)-(iv) of Assumption 2.1. As usual, the only part left is identifying an upper bound for $\bar{H}(x,\alpha,t)$. It is simple to prove under (7.8) that there is $M(\delta,\Delta)$ such that for each fixed $\delta > 0$, $M(\delta,\Delta)\to\infty$ as $\Delta\to 0$, and such that (for any N and $s\in[0,T]$)

$$P\{|X^N(s+\tau)-X^N(s)| \geqq \delta \text{ for some } \tau\in[0,\Delta]\} \leqq \exp -M(\delta,\Delta)/a_N. \tag{7.9}$$

Define $H(x,\alpha) = \log E\exp\langle\alpha, b_i(x)\rangle$. Using (7.9) and Theorem 7.2 (as Theorem 5.2 was used in the proof of Theorem 5.3), we obtain (with $c(N,t,\Delta) = m(t_N+t+\Delta)-m(t_N+t)$)

$$\begin{aligned}\bar{H}(x,\alpha,t) \leqq K^{-1}(t)\ &\overline{\lim_{\Delta\to 0}}\ \overline{\lim_{N\to\infty}} \log\Big(\exp\left[(H(x,K(t)\alpha)+O(\delta))c(N,t,\Delta)\right] \\ &+\exp\left[-M(\delta,\Delta)a_N^{-1}+\sup_y H(y,K(t)\alpha)c(N,t,\Delta)\right]\Big)\Big/ c(N,t,\Delta)\end{aligned}$$

where $O(\delta)\to 0$ as $\delta\to 0$. By Assumption 5.1, $a_N^{-1} \geqq \Delta c(N,t,\Delta)\theta$, for some $\theta > 0$. Taking limits in the order $\Delta\to 0$, then $\delta\to 0$, we obtain

$$\bar{H}(x,\alpha,t) \leqq K^{-1}(t)H(x,K(t)\alpha).$$

Example 7.2. Assume $\{\xi_n\}$ is generated as in Example 2.1, and define $F_n = b(X_n,\xi_n)$. Assume also that $b(\cdot,\cdot)$ is continuous and that it satisfies a linear growth condition in ξ, uniformly in x. Then Assumption 2.1^e holds. The sets Ω_N needed in defining $\bar{H}$ and in (7.1) are easily characterized in terms of the values of $|\xi_N|$ (see the remarks after the proofs of Theorems 4.1 and 7.1).

The simplest method of proving the α-differentiability of $\bar{H}(x,\alpha,t)$ at $\alpha = 0$ is to combine the method used in Example 7.1 with that of § 5.4, where in place of Assumption 5.2 we use the weaker assumption (Assumption 5.2^e) that requires uniformity only for $\omega\in\Omega_N$.

8. Concluding remarks and comparison with other results. There is a great deal of overlap in the models covered by our approach and those in [3], [8], [28], and [29]. All the methods are quite powerful. The method in [8] is similar (and seemingly simpler) than that in [28] and [29]. It allows for nonstationary state-dependent noise

and nonstationary correlated non-Markov noise, but it is restricted to $\sum a_i^2 < \infty$ and uses faster decrease conditions on the "tails" of the noise distributions. The method in [3] is simple, but the results for state-dependent noise or discontinuous dynamics are not very strong. The works [28] and [29] allow for slower decrease in the "tails" of the noise distributions than we do here, and for a slower rate of decrease in the correlation of the (Markov) noise, but are more restrictive in the requirements on $\{a_n\}$. They essentially require $\Sigma a_n^{1+\gamma} < \infty$ for some $\gamma > 0$ and that the sequence $\{a_n\}$ is ultimately decreasing. The latter condition is rather restrictive. They do not treat the constrained case. In fact, the method used in [28] and [29], as set up now, essentially requires stationarity in the noise process—or time homogeneity in the transition function for the state-dependent noise case (due to the method of construction and use of the solution to the Poisson equation defined in [29, p. 220]). Our upper-bounding technique allows for considerable nonstationarity. For example, we may consider processes driven by the noise model $\xi_{n+1} = A\xi_n + B\psi_n$, where the roots of A are inside the unit circle, the (bounded or Gaussian) ψ_n are mutually independent, zero mean, and where the covariances Σ_n are simply bounded. The differences will probably narrow as more work is done on both approaches.

Although we have not put the details in this paper, all the noise models can allow a more flexible x-dependence. For example,

$$P\{\xi_{n+1} \in A \mid X_i, i \leqq n+1, \xi_i, i \leqq n\} = P_0(X_{n+1}, X_n, X_{n-1}, \cdots, X_{n-k}, \xi_n, A)$$

where P_0 is continuous in $(X_{n+1}, \cdots, X_{n-k})$, uniformly in ξ_n, A. Similarly for models

$$\xi_n = \sum_{-\infty}^{n} c_{n-i}(X_n, \cdots, X_{n-k})\psi_{n-i}$$

with appropriate rate of decrease and continuity properties of the $\{c_n(\cdot)\}$. A further extension, not currently covered under the scheme in [28] and [29] (although that method can probably be extended to cover this case) occurs when all the components of F_n are not available simultaneously (e.g., as in adaptive data networks where time delays in the transmission of messages must be taken into account). For example, let the updating of the scalar components of X_n alternate.

Another feature of our approach is that we do not require continuity of the "mean" dynamics. (Note that the mean dynamics for the constrained case might be discontinuous just because of the boundary.)

Finally, we obtain a very useful upper bound to the "tails" of the escape probabilities, and have provided a setup where more general noise models can be treated—simply by appealing to new developments in the theory of large deviations. The use of the "continuity" method to handle the constrained problem (in § 6) illustrates how estimates for one (relatively simple) process can be converted into estimates for a much more complex process—simply by finding the appropriate "continuity map."

REFERENCES

[1] M. B. NEVELSON AND R. Z. HASMINSKII, *Stochastic Approximation and Recursive Estimation*, Translations of Mathematical Mongraphs, Vol. 47, American Mathematical Society, Providence, RI, 1972.

[2] M. T. WASAN, *Stochastic Approximation*, Cambridge University Press, London, 1969.

[3] H. J. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, 1978.

[4] M. METIVIER AND P. PRIOURET, *Applications of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, 30 (1984), pp. 140–151.

[5] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[6] H. J. KUSHNER AND H. HUANG, *Asymptotic properties of stochastic approximations with constant coefficients*, SIAM J. Control Optim., 19 (1981), pp. 87-105.

[7] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551-575.

[8] H. J. KUSHNER, *An averaging method for stochastic approximations with discontinuous dynamics, constraints, and state dependent noise*, in Recent Advances in Statistics, Rizvi, Rustagi, and Siegmund, eds., Academic Press, New York, 1983.

[9] A. P. KOROSTELEV, *Stochastic Recurrent Processes*, Nauka, Moscow, 1984. (In Russian.)

[10] H. J. KUSHNER, *Asymptotic behavior of stochastic approximation and large deviations*, IEEE Trans. Automat. Control, 29 (1984), pp. 984-990.

[11] P. DUPUIS AND H. J. KUSHNER, *Stochastic approximations via large deviations: asymptotic properties*, SIAM J. Control Optim., 23 (1985), pp. 675-696.

[12] M. I. FREIDLIN, *The averaging principle and theorems on large deviations*, Russian Math. Surveys, 33 (1978), pp. 117-176.

[13] H. J. KUSHNER, *Robustness and approximation of escape times and large deviations estimates for systems with small noise effects*, SIAM J. Appl. Math., 44 (1984), pp. 160-182.

[14] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, New York, 1979.

[15] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, 1984.

[16] J. GÄRTNER, *On large deviations from the invariant measure*, Theory Probab. Appl., 22 (1977), pp. 24-39.

[17] Y. SAISHO, *Stochastic differential equations for multi-dimensional domain with reflecting boundary*, Probab. Theory Rel. Fields, 74 (1987), pp. 455-477.

[18] H. J. KUSHNER AND P. DUPUIS, *Constrained stochastic approximation via the theory of large deviations*, in Adaptive Statistical Procedures and Related Topics, J. Van Ryzin, ed., IMS Lecture Notes—Monograph Series, Vol. 8, Institute of Mathematical Statistics, Hayward, CA, 1986.

[19] P. DUPUIS AND H. J. KUSHNER, *Asymptotic behavior of constrained stochastic approximations via the theory of large deviations*, Probab. Theory Rel. Fields, 75 (1987), pp. 223-244.

[20] P. DUPUIS, *Large deviations analysis of some recursive algorithms with state-dependent noise*, Ann. Probab., 16 (1988), pp. 1509-1536.

[21] S. R. S. VARADHAN, *Large Deviations and Applications*, CBMS-NSF Regional Conference Series in Applied Mathematics 46, Society for Industrial and Applied Mathematics, Philadelphia PA, 1984.

[22] S. OREY AND S. PELIKAN, *Large deviations for stationary processes*, IMA Preprint Series #227, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, 1986.

[23a] M. D. DONSKER AND S. R. S. VARADHAN, *Asymptotic evaluation of certain Markov process expectations for large time, Part I*, Comm. Pure Appl. Math., 28 (1975), pp. 1-47.

[23b] ———, *Asymptotic evaluation of certain Markov process expectations for large time, Part* II, 28 (1975), pp. 279-301.

[23c] ———, *Asymptotic evaluation of certain Markov process expectations for large time, Part* III, 29 (1977), pp. 389-461.

[23d] ———, *Asymptotic evaluation of certain Markov process expectations for large time, Part* IV, 36 (1983), pp. 183-212.

[24] I. ISCOE, P. NEY, AND E. NUMMELIN, *Large deviations of uniformly recurrent Markov additive processes*, Adv. in Appl. Math., 6 (1985), pp. 373-412.

[25] H. CRAMER, *Sur un nouveau théorème-limit de la théorie des probabilités*, in Colloque consacré à la théorie des probabilités, Vol. 3, Hermann, Paris, 1938.

[26] P. DUPUIS, *Large deviations analysis of reflected diffusions and constrained stochastic approximation algorithms in convex sets*, Stochastics, 21 (1987), pp. 63-96.

[27] R. AZENCOTT AND G. RUGET, *Mélanges d'équations différentielles et grands écarts à la loi des grands nombres*, Z. Wahrsch. Verw. Gebiete, 38 (1977), pp. 1-54.

[28] M. METIVIER AND P. PRIOURET, *Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant*, Probab. Theory Rel. Fields, 74 (1987), pp. 403-428.

[29] A. BENVENISTE, M. METIVIER, AND P. PRIOURET, *Algorithmes Adaptifs et Approximations Stochastiques*, Masson, Paris, 1987.

[30] H. KUSHNER, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: global minimization via Monte Carlo*, SIAM J. Appl. Math., 47 (1987), pp. 169-185.

[31] H. J. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, SIAM J. Control. Optim., 22 (1984), pp. 13-27.

[32] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1136-1155, September 1989

013

CONVEX DUALITY APPROACH TO THE OPTIMAL CONTROL OF DIFFUSIONS*

WENDELL H. FLEMING† AND DOMOKOS VERMES‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. An alternative to the usual dynamic programming approach to the optimal control of Markov processes is considered. It is based on duality of convex analysis. The control problem is embedded in a convex mathematical programming problem on a space of measures. The dual problem is to find the supremum of smooth subsolutions to the Hamilton-Jacobi-Bellman equation.

Key words. stochastic control, diffusion processes, convex duality, weak and strong solutions

AMS(MOS) subject classification. 93E20

1. Introduction. We consider $\mathbf{R}^n$-valued diffusion processes governed by the stochastic differential equation

$$dx_s = b(s, x_s, u_s)\,ds + \sigma(s, x_s, u_s)\,dw_s, \qquad x_t = x \tag{1.1}$$

with w_s an $\mathbf{R}^n$-valued Brownian motion and u_s a nonanticipative $Y \subset\subset \mathbf{R}^n$-valued control process. The objective is to minimize the expected (possibly discounted) cost

$$J^u(t, x) := \mathsf{E}^u_{t,x} \int_t^T e^{-c(s,x_s,u_s)} l(s, x_s, u_s)\,ds \tag{1.2}$$

over all control processes u. Here T is a finite or infinite planning horizon. Additional terminal costs could also be included.

An important feature of the present paper is that we do not make any ellipticity assumption; the matrix σ can be degenerate or even identically zero. This means the approach covers both deterministic and stochastic control theory.

Another specialty is that the running cost (and terminal cost if present) is not required to be bounded or continuous, but merely lower semicontinuous and of polynomial growth. This makes it possible, among other things, to also include problems where the objective is, e.g., to minimize the probability of the event that the state ever leaves a closed subset of the state space or to maximize the hitting probability of a target set; and in particular to cover the fixed endpoint problem of deterministic control theory.

In distinction from most papers in the field, the present approach does not use dynamic programming but is based on duality of convex analysis. We embed our control problem into a convex mathematical programming problem on a space of measures and consider its dual that turns out to involve the Hamilton-Jacobi-Bellman (HJB) equation. More precisely, we find that the dual of the original minimization

* Received by the editors January 25, 1988; accepted for publication (in revised form) January 5, 1989.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. This research was supported in part by the Institute for Mathematics and its Applications, National Science Foundation grant MCS-8121940, Office of Naval Research contract N00014-83-K0542, and by Air Force Office of Scientific Research contracts F-49620-86-C-0111 and AFOSR-85-0315.

‡ Department of Applied Mathematics, University of Washington, Seattle, Washington 98195. (On leave from the University of Szeged.) This research began while the author visited the Lefschetz Center for Dynamical Systems and was later supported by National Science Foundation grant DMS-8701768.

problem is to seek the supremum of all smooth subsolutions of the Hamilton-Jacobi-Bellman equation. From the existence of an equilibrium value for the primal-dual game it then follows, in particular, that the optimal value function of the control problem is the upper envelope of the smooth subsolutions of the Hamilton-Jacobi-Bellman equation.

The proof consists of two major steps. First we construct the minimization problem on the space of measures that contains the original control problem embedded (§ 3) and apply the Fenchel-Rockafellar duality theorem [4] to arrive at the HJB equation (§ 4). In the second step we prove that the embedding is actually tight; the infimum is the same both in the original and in the extended problem (§§ 5-6). This second part of the proof is based on the separation theorem and uses some analytic tools like mollification and Sobolev estimates, that in turn are derived by control-theoretic arguments. Roughly we could say that the separation is carried out by a sufficiently smooth control problem.

The possibility of approaching control problems via duality theory in abstract spaces was first demonstrated by Vinter and Lewis [6], [7] who proved similar results for deterministic control problems. Their approach was made available for stochastic control problems in [5] by basing it on the theory of occupation (potential and harmonic) measures and infinitesimal operators. The present paper extends the method to the optimal control of diffusions. Since the diffusion matrix is allowed to degenerate, the presented results apply uniformly to both deterministic and stochastic control problems. The novel proof of the tightness of the embedding is not only more general but even in the classical deterministic case it is more direct than the arguments of [6].

The first application of duality theory in optimal control was the representation of the adjoint co-state processes in terms of conjugate convex functions by Rockafellar [9]. A similar approach to stochastic control problems is due to Bismut [8]. In both cases the duality relationship is established in the finite-dimensional state/co-state spaces rather than between the spaces of measures and of continuous functions. Consequently both their methods and the nature of their results are very different from ours.

In [3] Lions characterizes the optimal value function of stochastic control as the largest generalized subsolution of the Hamilton-Jacobi-Bellman equation. The approach and method of proof differs from the one followed here.

2. Formulation of the problem. Let T be the planning horizon, either a nonnegative number or $+\infty$. We take $0 \leqq t \leqq T$. If $T < \infty$, then the state space will be $E^0 := [0, T] \times \mathbf{R}^n$, and if $T = +\infty$, then $E^0 := [0, T) \times \mathbf{R}^n$. We denote by E the one-point compactification of E^0 and introduce the notation $S^0 := E^0 \times Y$ and $S := E \times Y$. Note that E and S are compact.

The coefficients $\sigma(t, x, y)$ and $b(t, x, y)$ as well as the discount rate $c(t, x, y) \geqq 0$ are assumed to be bounded continuous functions on S^0 such that their first partial derivatives with respect to t and second partial derivatives with respect to x exist and, together with the functions themselves can continuously be extended to S. The running cost l is assumed to be lower semicontinuous on S and of at most polynomial growth. The case of additional terminal costs will be considered in § 8.

For simplicity we assume that either the planning horizon T is finite or that there is a strict discounting, i.e., $c_0 = \inf_{\sigma \in S^0} c(\sigma) > 0$. The effect of the discounting will be included into the process as an exponential killing or a jump to the fictitious isolated cemetery state Δ at the killing time Θ. In what follows all expectation signs E will refer to the killed process. The only exception is the sans serif E in formula (1.2) that

denotes the expectation of the nonkilled process, i.e.,

$$E\Phi(x_r) = \mathsf{E}\Phi(x_r) \cdot 1_{\{\Theta \geqq r\}} = \mathsf{E}\Phi(x_r)\, e^{-\int_t^r c(s,x_s,u_s)\, ds}.$$

We will also use the notation $\tau := \min(\Theta, T)$ and refer to it as the lifetime of the processes. The cost J^u can then be expressed in the three equivalent forms:

$$\begin{aligned} J^u(t,x) &= \mathsf{E}^u_{t,x} \int_t^T e^{-c(s,x_s,u_s)} l(s, x_s, u_s)\, ds \\ &= \mathsf{E}^u_{t,x} \int_t^\tau l(s, x_s, u_s)\, ds = E^u_{t,x} \int_t^T l(s, x_s, u_s)\, ds. \end{aligned} \tag{1.2'}$$

The assumptions about the boundedness of the coefficients, growth of the costs, and boundedness of the expected lifetime can be substantially relaxed. In fact, the proofs use a much less stringent but also less explicit assumption (cf. the remark following Lemma 2.1).

The spaces of functions on S^0 and E^0 that are continuously extendable to S and E will be denoted by $C(S)$ and $C(E)$, respectively, and they are considered to be Banach spaces normed by the supremum norm. In Lemma 2.1 we will introduce a continuous positive weight function $\gamma : [0, T) \times \mathbf{R}^n \to (0, \infty)$ associated with the control problem under investigation. We will consider the weighted spaces

$$C_\gamma(S) := \{f \in C(S^0) : f/\gamma \in C(S),\ \|f\|_\gamma := \sup_{\xi \in E, y \in Y} |f(\xi, y)|/\gamma(\xi) < \infty$$

$$\text{and } \lim_{|\xi| \to \infty} |f(\xi, y)|/\gamma(\xi) = 0\}.$$

$C_\gamma(E)$ is defined analogously.

$$C^2_\gamma(E) := \{\Phi \in C_\gamma(E) : \Phi(T,x)/\gamma(T,x) = 0,\ \Phi_t, \Phi_{x_i}, \Phi_{x_i x_j} \in C_\gamma(E) \quad \forall i, j = 1, \cdots, n\}.$$

In the subsequent expositions C^2_γ can always be substituted by the set of all infinitely often differentiable functions satisfying the boundary condition $\Phi(T,x)/\gamma(T,x) = 0$ and with all derivatives in $C_\gamma(E)$. We will refer to the elements of C^2_γ as smooth functions.

$\mathcal{M}^\gamma_\pm(S)$ will denote the space of all signed Borel measures M on S^0 for which the norm $\|M\|_\gamma = \int \gamma\, dM^+ + \int \gamma\, dM^-$ is finite. Here M^+ and M^- are the positive and negative parts of the Jordan decomposition of M. With obvious identification elements of $\mathcal{M}^\gamma_\pm(S)$ can be considered as signed measures on S not assigning mass to $\{\infty\} \times Y$.

If Γ is a positive constant then $\mathcal{M}^{\gamma,\Gamma}(S)$ will denote those nonnegative measures from $\mathcal{M}^\gamma_\pm(S)$ for which $\|M\|_\gamma \leqq \Gamma < +\infty$.

The set $\mathcal{U}$ of all admissible controls consists of all Y-valued control processes u_s which are progressively measurable with respect to the filtration of the Brownian motion w_s. If $u \in \mathcal{U}$ then x^u_s denotes the solution of the stochastic differential equation (1.1) corresponding to u, satisfying the initial condition $x^u_t = x$ and killed at rate $c(\cdot)$. The corresponding expectation operator will be denoted by $E^u_{t,x}$ and if no confusion can arise, the superscript u will be omitted from x^u_t inside the expectation.

With each control $u \in \mathcal{U}$ we associate the measure M^u defined on the compact space $S = E \times Y$ which is the extension of

$$\begin{aligned} M^u(B_t \times B_x \times B_y) &:= E^u_{t,x} \int_{[t,T] \cap B_t} 1_{B_x}(x^u_s) \cdot 1_{B_y}(u_s)\, ds \\ M(\infty \times Y) &:= 0. \end{aligned} \tag{2.1}$$

Here $B_t \subset [0, \infty]$, $B_x \subset \mathbf{R}^n$, $B_y \subset Y$ are arbitrary Borel sets and 1_B denotes the indicator function of the set B. Note that though the notation does not indicate it, the measures M^u depend on the initial condition $x_t = x$ in (1.1) which is considered to be fixed. We will denote the set of all such M^u corresponding to some $u \in \mathcal{U}$ by $\mathfrak{M}^S(t, x)$.

Intuitively, $M([t, t'] \times B_x \times B_y)$ measures the expected time before t' spent by the *killed* process x_s^u in the set B_x while control values from $B_y \subset Y$ were applied. In particular, $M^u(\cdot, \cdot, Y)$ is the potential (or occupation) measure of the killed time-space process (s, x_s^u).

The infinitesimal operator of the killed Markov process x_t^y corresponding to the constant control $u_t \equiv y \in Y$ is defined for each $\Phi \in C^2(E^0)$ and is given by the expression

$$A^y\Phi(t, x) = \frac{\partial \Phi(t, x)}{\partial t} + \sum_{i,j=1}^{n} a_{i,j}(t, x, y) \frac{\partial^2 \Phi(t, x)}{\partial x_i \, \partial x_j} + \sum_{i=1}^{m} b_i(t, x, y) \frac{\partial \Phi(t, x)}{\partial x_i}$$

$$- c(t, x, y)\Phi(t, x)$$

with $(a_{ij}) = \frac{1}{2}\sigma^T \cdot \sigma$. We will use this notation also for nonsmooth functions Φ i.e., to denote the value of the expression on the right-hand side at every point (t, x) where the corresponding partial derivatives exist.

To interconnect the assumptions on discounting, termination and growth as well as to express them in a technically convenient analytic form, we prove the following lemma.

LEMMA 2.1. *There exist constants $0 < \alpha < \bar{\alpha}$ and a twice continuously differentiable function $\gamma\colon [0, T) \times \mathbf{R}^n \to (0, \infty)$ satisfying*

$$0 < \alpha\gamma \leqq -A^y\gamma \leqq \bar{\alpha}\gamma \tag{2.2}$$

everywhere in $(0, T) \times \mathbf{R}^n$ for all $y \in Y$.

Proof. We will construct γ separately for the discounted and for the finite horizon case.

1. *Discounted case.* The infinitesimal operator of the exponentially killed process is of the form $A\Phi = D\Phi - c\Phi$ with D a (possibly degenerate) second order differential operator. We define

$$\gamma(t, x) := (\cosh pt) \prod_{i=1}^{n} \cosh px_i \tag{2.3}$$

with a p yet to be determined.

A straightforward calculation shows that

$$-K(p)\gamma(t, x) \leqq D\gamma(t, x) \leqq K(p)\gamma(t, x) \tag{2.4}$$

with $K(p) = \sum_{i=1}^{n} \|a_{ij}\| p^2 + \sum_{i=1}^{n} \|b_i\| p = \bar{a}p^2 + \bar{b}p$. Consequently,

$$K(p) + \|c\| \cdot \gamma \leqq A\gamma = D\gamma - c \cdot \gamma \leqq (K(p) - c) \cdot \gamma. \tag{2.5}$$

If $c_0 = \inf c > 0$, then the quadratic equation $\bar{a}p^2 + \bar{b}p - c_0 = 0$ has exactly one positive root p_0. Choosing p from the interval $(0, p_0)$, we get that $c_0 - K(p) > 0$, and hence (2.2) is satisfied with $\alpha := c_0 - K(p)$ and $\bar{\alpha} := \|c\| + K(p)$.

2. *Finite horizon case.* We define

$$\gamma(t, x) := [1 + (T - t)] \cdot \prod_{i=1}^{n} \cosh px_i = [1 + (T - t)]\gamma_0(x).$$

When we use the notation $A\Phi = \partial\Phi/\partial t + D_x\Phi$ a calculation analogous to that of the discounted case yields

$$
\begin{aligned}
-A\gamma(t,x) &= \gamma_0(x) - (1+T-t)D_x\gamma_0(x) \\
&\geqq [(1+T-1)^{-1} - K(p)] \cdot (1+T-t) \cdot \gamma_0(x).
\end{aligned} \tag{2.6}
$$

With $(1+T)^{-1}$ in place of c_0, the above argument shows that if p is chosen from $(0, p_0)$, then γ satisfies (2.2) with $\alpha := (1+T)^{-1} - K(p) > 0$ and $\bar{\alpha} := 1 + K(p)$. The proof of the lemma is complete.

We formulate some consequences of Lemma 2.1 that will be used at various places during the subsequent expositions.

COROLLARY. (1) $\int \gamma\, dM^u \leqq \gamma(t,x)/\alpha < +\infty$ *for every* $M^u \in \mathfrak{M}^S(t,x)$. *In other words, the constant* $\Gamma := \gamma(t,x)/\alpha < +\infty$ *is a uniform upper bound for the expressions* $E^u_{t,x}\int_t^T \gamma(s, x^u_s)\, ds$ *for every process* x^u_t *generated by a control* $u \in \mathcal{U}$ *and starting from initial state* $x_t = x$.

(2) $\gamma(t,x)$ *grows asymptotically not faster than an exponential function as* $|x| \to \infty$, $t \to \infty$.

(3) *For every* $(t,x) \in E^0$, $0 \leqq s < +\infty$ *and* $u \in \mathcal{U}$ *we have*

$$
1 - e^{-\alpha s} \leqq 1 - \gamma^{-1}(t,x)E^u_{t,x}\gamma(t+s, x_{t+s}) \leqq 1. \tag{2.7}
$$

Proof. The proof of (1) follows from Dynkin's formula. In fact, if $T < \infty$ we have

$$
\begin{aligned}
\int \gamma\, dM^u &\leqq \frac{1}{\alpha}\int (-A^u\gamma)\, dM^u = \frac{1}{\alpha} E^u_{t,x}\int_t^T (-A^u\gamma)(s, x^u_s)\, ds \\
&= \frac{1}{\alpha}[\gamma(t,x) - E^u_{t,x}\gamma(T, x_T)] \leqq \gamma(t,x)/\alpha.
\end{aligned}
$$

Since the bound is independent of T, the inequality remains true as $T \to +\infty$.

The proof of (2) is immediate from the construction of γ in the proof of Lemma 2.1.

(3) The left-hand side of (2.2) can be written as $A^u\gamma + \alpha\gamma \leqq 0$. By the Feynman–Kac formula it follows that $E^u_{t,x}e^{\alpha s}\gamma(t+s, x^u_{t+s}) \leqq \gamma(t,x)$ with an $\alpha > 0$. Subtracting both sides of inequality $E^u_{t,x}\gamma(t+s, x^u_{t+s}) \leqq e^{-\alpha s}\gamma(t,x)$ from $\gamma(t,x)$, we obtain $\gamma - E\gamma \geqq \gamma(1 - e^{-\gamma s})$, which proves the left-hand side of (2.7). The right-hand side is trivial since $\gamma^{-1}E\gamma \geqq 0$.

Remark. The growth, discounting, and termination conditions required earlier in this section will be used in the subsequent expositions only indirectly through the statement of Lemma 2.1. Consequently all results of this paper remain valid under other sets of assumptions that assure the existence of a γ with property (2.2). Examples of other possible sets of such assumptions are:

(i) Coefficients a_{ij}, b_i satisfy linear growth conditions, the discounting is strict, the running cost is bounded. In this case γ can be chosen asymptotically as $|x|^p$ with $p < c_0$ and $\alpha = c_0 - p$.

(ii) Coefficients a_{ij}, b_i satisfy linear growth conditions, the time horizon is finite, the running cost is of polynomial growth. Then we can choose $\gamma(t,x) \sim [1 + K(T-t)]|x|^p$ with an appropriate K and p.

Now we return to our original control problem. Although we assumed l to be only lower semicontinuous, in §§ 3–6 of the paper we will consider continuous running costs. The extension of all obtained results to the general semicontinuous case will be an additional step in § 7. With the notation introduced, the control problem we will consider in §§ 3–6 can be formulated as the Strong Problem.

STRONG PROBLEM. For a given running cost $l \in C_\gamma(E)$ and initial state $(t, x) \in E^0$:

$$\text{Minimize} \int l\, dM^u \quad \text{over all } M^u \in \mathfrak{M}^S(t, x).$$

We can define the optimal value ψ of the Strong Problem as a function of the initial state

$$\psi(t, x) := \inf\left\{ \int l\, dM^u : M^u \in \mathfrak{M}^S(t, x) \right\}.$$

3. The weak formulation of the control problem. It follows from Itô's formula, that for arbitrary nonanticipative control process $u \in \mathcal{U}$ the generalization of the fundamental theorem of calculus (Dynkin's formula) holds true. For every twice continuously differentiable Φ we have

$$E^u_{t,x}\Phi(\sigma, x_\sigma) - \Phi(t, x) = E^u_{t,x} \int_t^\sigma A^{u_s}\Phi(s, x_s)\, ds \tag{3.1}$$

provided $\sigma \leqq \tau$ is a stopping time such that the expectations exist.

If we apply this formula to the terminal time τ and to smooth functions $\Phi \in C^2_\gamma$ that vanish at the terminal state Δ, then by $\Phi(\tau, x_\tau) = \Phi(\Delta) = 0$ we find that

$$-\Phi(t, x) = \int A^\gamma \Phi(t', x') M^u(dt', dx', dy) \tag{3.2}$$

holds true for every $u \in \mathcal{U}$ whenever $A\Phi \in C_\gamma$.

We introduce the notation

$$\mathcal{M}_A(t, x) := \left\{ M \in \mathcal{M}^\gamma_\pm(S) : -\Phi(t, x) = \int A\Phi\, dM \quad \text{for all } \Phi \in C^2_\gamma(E) \right\},$$

$$\mathfrak{M}^W(t, x) := \mathcal{M}^{\gamma,\Gamma}(S) \cap \mathcal{M}_A(t, x) \quad \text{with } \Gamma = \gamma(t, x)/\alpha.$$

Since for every $u \in \mathcal{U}$ the measure $M^u \in \mathfrak{M}^S(t, x)$ is in both $\mathcal{M}^{\gamma,\Gamma}(S)$ and $\mathcal{M}_A(t, x)$ our original control problem, the Strong Problem is embedded in the following problem.

WEAK PROBLEM.

$$\text{Minimize} \int l\, dM \quad \text{over } M \in \mathfrak{M}^W(t, x). \tag{3.3}$$

This is a minimization problem on the space of measures with linear objective and convex constraints. In fact, if $l \in C_\gamma(S)$, then by the Riesz Theorem $\int l\, dM$ is a continuous linear functional on the space of signed measures $\mathcal{M}^\gamma_\pm$. For each $\Phi \in C^2_\gamma$ relation (3.2) imposes a continuous linear restriction on M, consequently their intersection $\mathcal{M}_A(t, x)$ is a closed linear set in $\mathcal{M}^\gamma_\pm$. Finally, $\mathcal{M}^{\gamma,\Gamma}$ is a w^*-compact convex subset of $\mathcal{M}^\gamma_\pm$.

The feasible set of the Strong Problem consists of all $M^u \in \mathfrak{M}^S$ generated by a control $u \in \mathcal{U}$ via the stochastic differential equation (1.1). This set is contained in the feasible set $\mathfrak{M}^W$ of the weak problem, thus the optimal value $\psi(t, x) := \inf\{\int l\, dM^u : u \in \mathcal{U}\}$ is not less than the minimum $\Psi(t, x) = \inf\{\int l\, dM;\ M \in \mathfrak{M}^W(t, x)\}$ in the Weak Problem. Note that the initial state (t, x) is involved in the Strong Problem through the initial condition (1.2) and in the Weak Problem through the definition of $\mathcal{M}_A(t, x)$.

In what follows, we will first characterize the value function $\Psi(t, x)$ of the Weak Problem by solving its dual, a maximization problem in the function space $C_\gamma(E) \subset C_\gamma(S)$. More precisely, it will turn out that the dual of the minimization problem (3.3) is to find the supremum of all smooth subsolutions to the Hamilton-Jacobi equation.

To make duality methods applicable it is convenient to bring the Weak Problem to the Fenchel normal form. Using extended valued functions we reformulate the convexly constrained linear problem as an unconstrained convex problem. In fact, we introduce the functionals h_1 and $h_2: \mathcal{M}_+^\gamma(S) \to \bar{\mathbf{R}}^1$ by

$$h_1(M) := \begin{cases} \int l\, dM & \text{if } M \in \mathcal{M}^{\gamma,\Gamma}(S), \\ +\infty & \text{otherwise,} \end{cases}$$

$$h_2(M) := \begin{cases} 0 & \text{if } M \in \mathcal{M}_A(t, x), \\ -\infty & \text{otherwise.} \end{cases}$$

Both h_1 and $-h_2$ are convex and lower semicontinuous. It is immediate that the weak problem is equivalent to the following problem.

FENCHEL PROBLEM. Minimize $h_1(M) - h_2(M)$ over all $M \in \mathcal{M}_+^\gamma(S)$.

4. Duality and the Hamilton–Jacobi problem. Recall that the space S is compact, thus by the Riesz Theorem $C_\gamma^*(S) = \mathcal{M}_+^\gamma(S)$. In other words, $C_\gamma(S)$ and $\mathcal{M}_+^\gamma(S)$ are spaces in duality connected by the bilinear form:

$$\langle \phi, \mu \rangle = \int \phi\, d\mu \qquad \phi \in C_\gamma, \quad \mu \in \mathcal{M}_+^\gamma. \tag{4.1}$$

The norm topology of C_γ and the weak*-topology of $\mathcal{M}_+^\gamma$ are compatible with the pairing, the continuous linear functionals on both spaces are exactly those representable by the bilinear form. If H and h are convex real-valued functions defined on $C_\gamma(S)$ and $\mathcal{M}_+^\gamma(S)$, respectively, then their Legendre-Fenchel transforms (convex conjugates) are defined by

$$H^*(\mu) := \sup \left\{ \int \phi\, d\mu - H(\phi) \colon \phi \in C_\gamma(S) \right\}, \tag{4.2}$$

$$h^*(\phi) := \sup \left\{ \int \phi\, d\mu - h(\mu) \colon \mu \in \mathcal{M}_+^\gamma(S) \right\}. \tag{4.3}$$

If the original function h or H is convex and lower semicontinuous, then it coincides with its double conjugate, i.e., $H^{**} = H$, $h^{**} = h$. Conjugates of concave functions are defined analogously but with inf in place of sup, and have the corresponding properties.

Now we compute the Legendre-Fenchel transforms of the functionals h_1 and h_2. We use the quantities γ and α as they were introduced in Lemma 2.1.

LEMMA 4.1. $h_1^*(\phi) = \alpha^{-1} \cdot \gamma(t, x) \cdot \|(\phi - l)^+\|_\gamma = \alpha^{-1} \cdot \gamma(t, x) \cdot \sup\{[\phi(\sigma) - l(\sigma)]/\gamma(\theta, \xi)$: *over all* $\sigma = (\theta, \xi, \eta) \in S$ *such that* $\phi(\sigma) - l(\sigma) \geqq 0\}$.[1]

Proof.

$$\begin{aligned} h_1^*(\phi) &= \sup \left\{ \int \phi\, d\mu - h_1(\mu) \colon \mu \in \mathcal{M}_+^\gamma \right\} = \sup \left\{ \int (\phi - l)\, dM \colon M \in \mathcal{M}^{\gamma,\Gamma} \right\} \\ &= \sup \left\{ \int [(\phi - l)/\gamma] \gamma\, dM \colon M \geqq 0, \int_\gamma dM \leqq \Gamma = \gamma(t, x)/\alpha \right\}. \end{aligned} \tag{4.4}$$

[1] $(f)^+$ denotes the positive part of the function f, i.e., $f^+(x) = \max\{0, f(x)\}$.

Since ϕ and l are in C_γ, the continuous function $(\phi - l)/\gamma$ attains its maximum at some point $\sigma_0 = (t_0, x_0, y_0)$ of the compact set S. If $(\phi - l)(t_0, x_0, y_0)/\gamma(t_0, x_0) > 0$, then $t_0 < \infty$, $x_0 \neq \infty$ and the sup in (3.4) can be attained by concentrating all available mass of the measure $\gamma\, dM$ to the point $\sigma_0 \in S$. We have to choose $M(ds) := \gamma(t, x)/(\alpha \cdot \gamma(t_0, x_0))\delta_{\sigma_0}(ds)$ with δ_{σ_0} denoting the Dirac measure assigning unit mass to the singleton $\{\sigma_0\}$. Then we have

$$h_1^*(\phi) = \gamma(t, x)(\phi - l)(\sigma_0)/(\alpha \cdot \gamma(t_0, x_0)) = \alpha^{-1} \cdot \gamma(t, x)\|(\phi - l)\|_\gamma \tag{4.5}$$

provided $\sup(\phi - l) > 0$.

If $\sup(\phi - l) \leqq 0$, i.e., if $\phi(\sigma) < l(\sigma)$ for all $\sigma \in S$, then the maximum of the expression (4.4) is zero and is attained for $M \equiv 0$. This together with (4.5) proves the lemma.

LEMMA 4.2.

$$h_2^*(\phi) = \begin{cases} -\lim \Phi_i(t, x) & \text{if } \phi = \lim_{i\to\infty} A\Phi_i \text{ with } \Phi_i \in C_\gamma^2(E), \\ -\infty & \text{otherwise.} \end{cases} \tag{4.6}$$

Proof. Since h_2 is concave, $h_2^*(\phi) := \inf\{\int \phi\, dM - h_2(M)\colon M \in \mathcal{M}_A(t, x)\}$.

Let us first assume that $\phi = A\Phi \in C_\gamma$ with some $\Phi \in C_\gamma^2$. Then, by the definition of $\mathcal{M}_A(t, x)$ for every $M \in \mathcal{M}_A(t, x)$ we have $\int \phi\, dM = \int A\Phi\, dM = -\Phi(t, x)$. Since $\mathcal{M}_A$ is nonempty, $A\Phi_1 \equiv A\Phi_2$ implies $\Phi_1(t, x) = \Phi_2(t, x)$, and hence we have $\int \phi\, dM = \int A\Phi\, dM = -\Phi(t, x)$ whenever $\phi = A\Phi \in C_\gamma$ with some $\Phi \in C_\gamma^2$.

Let us assume now that there exists a sequence $\Phi^k \in C_\gamma^2$ such that $\|\phi - A\Phi^k\|_\gamma \to 0$. This means that $A\Phi^k/\gamma \to \phi/\gamma$ uniformly on S as $k \to \infty$. By this uniform convergence and the finiteness of the measure $\gamma\, dM$ for every $M \in \mathcal{M}_A(t, x) \subset \mathcal{M}_\pm^\gamma$ we have

$$\int \phi\, dM = \int \frac{\phi}{\gamma} \cdot \gamma\, dM = \int \lim_{k\to\infty} \frac{A\Phi^k}{\gamma} \cdot \gamma\, dM = \int \lim_{k\to\infty} A\Phi^k\, dM = \lim \Phi^k(t, x)$$

independently of the particular choice of the sequence $A\Phi^k$. Since $-\lim \Phi^k(t, x)$ does not depend on $M \in \mathcal{M}_A(t, x)$, we have proved the first line of (4.6).

It remains to show that $h_2^*(\phi) = -\infty$ if ϕ is not in the $\|\ \|_\gamma$ closure of the functions $A\Phi$ with $\Phi \in C_\gamma^2$. Assume $\phi_0 \in C_\gamma$ is not in the closed subspace $W := \{\phi \in C_\gamma(S)\colon \lim_{k\to\infty} \|\phi - A\Phi^k\|_\gamma = 0 \text{ with } \Phi^k \in C_\gamma^2\}$. Then ϕ_0 and W can be separated by a closed hyperplane. That is, there exists an $M' \in \mathcal{M}_\pm^\gamma(S)$ such that $\int \phi_0\, dM' < 0$ while $\int \phi\, dM' = 0$ for all $\phi \in W$. In particular, we have $\int A\Phi\, dM' = 0$ for all $\Phi \in C_\gamma^2$ and consequently, $M + \Theta M' \in \mathcal{M}_A(t, x)$ for every $M \in \mathcal{M}_A(t, x)$ and $\Theta \in \mathbf{R}^1$. If $\bar{M}$ denotes an arbitrary fixed element of $\mathcal{M}_A(t, x)$, then we have

$$h_2^*(\phi_0) = \inf_{M\in\mathcal{M}_A} \int \phi_0\, dM \leqq \inf_{\Theta\in\mathbf{R}^1} \int \phi_0\, d(\bar{M} + \Theta M')$$

$$= \int \phi_0\, d\bar{M} + \inf_{\Theta\in\mathbf{R}^1} \Theta \cdot \int \phi_0\, dM' = -\infty.$$

Here we used that by assumption $\int \phi_0\, dM' \neq 0$ and that Φ can be arbitrary. This completes the proof of the lemma.

The next theorem is the main result of this section. Roughly, it states that seeking the maximal solution of the Hamilton-Jacobi-Bellman equation is the dual to the weak problem formulated in the previous section. As under the current weak assumptions no smooth solutions to the Hamilton-Jacobi-Bellman equation need exist, the precise formulation of the duality relationship is the following. The value function

(i.e., the minimum) of the Weak Problem is the upper envelope (i.e., supremum) of the smooth subsolutions of the Hamilton-Jacobi equation.

THEOREM 1.

$$\Psi(t,x) := \min\left\{\int l\,dM\colon M \in \mathcal{M}_A(t,x) \cap \mathcal{M}^{\gamma,\Gamma}\right\}$$

$$= \sup\{\Phi(t,x)\colon \Phi \in C_\gamma^2,\ A\Phi + l \geqq 0\}.$$

Proof. If applied to $C_\gamma^* = \mathcal{M}_\pm^\gamma$, Rockafellar's duality theorem [4] states that

$$\min\{h_1(M) - h_2(M)\colon M \in \mathcal{M}_\pm^\gamma(S)\} = \sup\{h_2^*(\phi) - h_1^*(\phi)\colon \phi \in C_\gamma(S)\} \tag{4.7}$$

whenever the set $\{\phi\colon h_2^*(\phi) > -\infty\}$ contains a finite continuity point of h_1^*. But this condition is satisfied since h_1^* is continuous and finite on whole C_γ and $h_2^*(\phi)$ is not identically $-\infty$, and hence (4.7) holds true.

Substituting the explicit expressions for h_1^* and h_2^* from Lemmas 4.1 and 4.2 into (4.7) and using the fact that $\{A\Phi\colon \Phi \in C_\gamma^2\}$ is dense in $\{\phi\colon h_2^*(\phi) > -\infty\}$, we obtain

$$\Psi(t,x) = \min\{h_1(M) - h_2(M)\colon M \in \mathcal{M}_\pm^\gamma(S)\}$$

$$= \sup\{\Phi(t,x) - \alpha^{-1}\cdot\gamma(t,x)\cdot\|(A\Phi + l)^-\|_\gamma\colon \Phi \in C_\gamma^2\}.$$

To conclude the proof it is sufficient to show that for every $\Phi \in C_\gamma^2$ there exists a $\Phi^\sim \in C_\gamma^2$ such that $A\Phi^\sim + l \geqq 0$ and $\Phi^\sim(t,x) \geqq \Phi(t,x) - \alpha^{-1}\cdot\gamma(t,x)\cdot\|(A\Phi + l)^-\|_\gamma$. Choose $\Phi^\sim := \Phi - \alpha^{-1}\cdot\gamma\|(A\Phi + l)^-\|_\gamma$. Then by Lemma 2.1 $-A\gamma \geqq \alpha\gamma$ holds and consequently, we have

$$A\Phi^\sim + l = A\Phi + l - \alpha^{-1}\cdot\|(A\Phi + l)^-\|_\gamma\cdot A\gamma \geqq A\Phi + l + \gamma\cdot\|(A\Phi + l)^-\|_\gamma$$

$$= A\Phi + l + \gamma\cdot\sup_{(t',x',y')\in S}|(A\Phi + l)^-(t',x',y')/\gamma(t',x')| \geqq 0.$$

The proof of the theorem is complete.

In a less compressed form Theorem 1 states that the weak value function Ψ is the upper envelope of all $\Phi \in C_\gamma^2(E)$ satisfying the Hamilton-Jacobi inequality:

$$\Phi_t(t,x) + \min_{y\in Y}\left\{\sum_{i,j=1}^n a(t,x,y)\Phi_{x_ix_j}(t,x) + \sum_{i=1}^n b_i(t,x,y)\Phi_{x_i}(t,x)\right.$$

$$\left. - c(t,x,y)\Phi(t,x) + l(t,x,y)\right\} \geqq 0. \tag{4.8}$$

Recall that the definition of C_γ^2 includes $\Phi(T,x) = 0$ whenever $T < +\infty$.

The fact that $A\Phi_1 \geqq A\Phi_2$ implies $\Phi_1 \leqq \Phi_2$ justifies calling the functions $\Phi \in C_\gamma^2$ satisying (4.8) *subsolutions* of the Hamilton-Jacobi equation.

The results of the present paragraph remain valid under much more general assumptions than those made in § 2. In fact, we did *not* use either the finite dimensionality of the state-space or the specific properties of diffusion processes. Besides Rockafellar's duality theorem, our approach was based on the validity of Dynkin's formula, but not even the denseness of C_γ^2 was exploited. Since Dynkin's formula is a special case of the "general fundamental theorem of calculus" in semigroup theory, all results of the present paragraph can be generalized to the case, when the state and control spaces are locally compact separable metric spaces and C_γ^2 is substituted by a linear subset $\mathcal{L}$ of $C_\gamma(E)$. Of course, this latter change affects the definition of $\mathcal{M}_A$ and consequently the Weak Problem itself. But still, the dual of this new "$\mathcal{L}$-Weak" Problem will be the problem of finding the upper envelope of all subsolutions in $\mathcal{L}$

of the Hamilton-Jacobi-Bellman equation involving the operator A. The coincidence of the primal and dual values remain preserved too.

5. Equivalence of the strong and weak formulations. We prove the equivalence under the assumption of a special approximation property of the value function corresponding to smooth costs. In § 6 we will show that under the assumptions of the present paper this approximability is always true.

THEOREM 2. *Let $f \in C_\gamma^2(S)$ denote an arbitrary smooth "running cost" and denote F the corresponding (strong) value function.*

Suppose that every such value function F can be approximated in the $\| \ \|_\gamma$-norm by a sequence of functions $F^{(\varepsilon)}$, each of which has first and second derivatives essentially bounded in γ-norm and satisfies $AF^{(\varepsilon)} + f \geqq 0$ almost everywhere as well as $F^{(\varepsilon)}(T, x) = 0$ whenever $T < \infty$. Then, for each $(t, x) \in E^0$ and l as in § 2 the weak and strong formulations are equivalent; their optimal value functions coincide.

Note that Theorem 2 assumes the approximability of value functions generated by smooth costs and makes a statement about the more general control problem that involves general continuous or (later) even only lower semicontinuous running cost l.

Proof. Assume that the statement of the theorem is false; then there exists an initial state (t_0, x_0) such that $\Psi(t_0, x_0) < \psi(t_0, x_0)$. This means that there exists a measure $M_0 \in \mathfrak{M}^W(t_0, x_0) \setminus \mathfrak{M}^S(t_0, x_0)$ that gives rise to a cost $\int l \, dM_0$ lower than $\psi(t_0, x_0)$, the infimum over all costs generated by controls $u \in \mathcal{U}$, i.e.,

$$\int l \, dM_0 < \inf \left\{ \int l \, dM^u : u \in \mathcal{U} \right\}. \tag{5.1}$$

This means that the w^*-continuous linear functional $\int l \, dM$ on $\mathcal{M}_+^\gamma(S)$ separates an element $M_0 \in \mathfrak{M}^W$ from the w^* convex-closure of the set $\mathfrak{M}^S = \{M^u : u \in \mathcal{U}\}$. In other words, $\mathfrak{M}^W$ is strictly larger than the closure of $\mathfrak{M}^S$. If this is so, then M_0 and the compact set $\bar{\mathfrak{M}}^S$ can also be separated by a functional $\int f \, dM$ generated by a smooth $f \in C_\gamma^2(S)$. More precisely, since smooth functions form a dense subset in C_γ there must exist an $f \in C_\gamma^2$ such that

$$\int f \, dM_0 < \inf \left\{ \int f \, dM^u : M^u \in \mathfrak{M}^S \right\}. \tag{5.2}$$

Let us introduce the strong value function F corresponding to the running cost f

$$F(t, x) := \inf \left\{ \int f \, dM : M \in \mathfrak{M}^S(t, x) \right\} = \inf_{u \in \mathcal{U}} E_{t,x} \int_t^T f(t, x_t^u, u_t) \, dt. \tag{5.3}$$

Then, according to the assumptions of the theorem, for every $\varepsilon > 0$ there exists an $F^{(\varepsilon)}$ such that all the partial derivatives $F_t^{(\varepsilon)}$, $F_{x_i}^{(\varepsilon)}$, $F_{x_i x_j}^{(\varepsilon)}$ are defined almost everywhere, are essentially bounded, and for every $y \in Y$ the inequality

$$A^y F^{(\varepsilon)}(t, x) + f(t, x, y) \geqq 0 \tag{5.4}$$

is satisfied for almost every $(t, x) \in E$ and $\| F^{(\varepsilon)} - F \|_\gamma < \varepsilon$.

The generalized Dynkin formula (3.2) cannot be applied directly to (5.4) because $F^{(\varepsilon)}$ is not smooth; it should first be approximated by C_γ^2 functions. The details of this approximation are presented in the next two lemmas. When we use them, the conclusion of the proof of Theorem 2 will be straightforward.

LEMMA 5.1. *For every $\delta > 0$ there exists an $F^{(\varepsilon,\delta)} \in C_\gamma^2(E)$ such that*

$$\begin{aligned} &\| F^{(\varepsilon)} - F^{(\varepsilon,\delta)} \|_\gamma < \delta, \qquad \| AF^{(\varepsilon,\delta)} \|_\gamma \leqq \| AF^{(\varepsilon)} \|_\gamma + \delta, \\ &AF^{(\varepsilon,\delta)} + f \geqq -\delta \cdot \gamma \quad \text{on } [\delta, T - \delta] \times \mathbf{R}^n \times Y. \end{aligned} \tag{5.5}$$

Proof. First we extend the definition of $F^{(\varepsilon)}$ from $[0, T]\times\mathbf{R}^n$ to $[-T, 2T]\times\mathbf{R}^n$ by

$$\begin{aligned} &F^{(\varepsilon)}(-s, x) := F^{(\varepsilon)}(0, x), \qquad s\in[0, T], \quad x\in\mathbf{R}^n, \\ &F^{(\varepsilon)}(T+s, x) := F^{(\varepsilon)}(T-s, x), \end{aligned} \tag{5.6}$$

and the functions a, b, c, and f from $[0, T]\times\mathbf{R}^n\times Y$ to $[-T, 2T]\times\mathbf{R}^n\times Y$ by reflection over 0 and T; i.e., $a(-s, x, y)=a(s, x, y)$ and $a(T+s, x, y)=a(T-s, x, y)$ if $s\in[0, T]$, $x\in\mathbf{R}^n$, $y\in Y$, and similarly for b, c, and f.

Note that

$$\begin{aligned} &A^y F^{(\varepsilon)}(-s, x) = -F_t^{(\varepsilon)}(s, x) + A^y F^{(\varepsilon)}(s, x), \qquad s\in[0, T], \quad x\in\mathbf{R}^n, \quad y\in Y, \\ &A^y F^{(\varepsilon)}(T+s, x) = 2F_t^{(\varepsilon)}(s, x) - A^y F^{(\varepsilon)}(T-s, x). \end{aligned} \tag{5.7}$$

Moreover, because of the Lipschitz continuity of $F^{(\varepsilon)}$ we have

$$\sup_{[-T,2T]\times\mathbf{R}^n\times Y} |AF^{(\varepsilon)}/\gamma| = K$$

with some finite number K. (We reserve the notation $\|\cdot\|_\gamma$ for sup over $[0, T]\times\mathbf{R}^n\times Y$.)

Let $\rho_r(t, x)$ be a nonnegative symmetric C^∞-mollifier (partition of unity) with $\iint \rho_r(\sigma, \xi)\, d\sigma\, d\xi = 1$ and $\rho_r(\sigma, \xi)=0$ if $|\sigma|+|\xi|>r$. If $\phi\in C([-T, 2T]\times\mathbf{R}^n)$, then we define $\phi * \rho_r$ on $[0, T]\times\mathbf{R}^n$ by

$$(\phi * \rho_r)(t, x) = \iint \phi(t+\sigma, x+\xi)\rho_r(\sigma, \xi)\, d\sigma\, d\xi \quad \text{if } 0<r<T.$$

From the second relation of (5.6) it follows that $(F^{(\varepsilon)} * \rho_r)(T, x)=0$; moreover, $F^{(\varepsilon)} * \rho_r$ is infinitely often differentiable on $[0, T]\times\mathbf{R}^n$ and $\|F^{(\varepsilon)} * \rho_r\|_\gamma \leqq K$. Consequently, $F^{(\varepsilon)} * \rho_r \in C_\gamma^2(E)$ for every $0<r<T$.

Since by (5.4) $AF^{(\varepsilon)}+f\geqq 0$ holds almost everywhere on $[0, T]\times\mathbf{R}^r\times Y$, it follows that

$$(AF^{(\varepsilon)}) * \rho_r + f * \rho_r \geqq 0 \quad \text{on } [r, T-r]\times\mathbf{R}^n\times Y.$$

We want to show that for every $\delta>0$ there exists an $r>0$ such that $F^{(\varepsilon,\delta)} := F^{(\varepsilon)} * \rho_r$ satisfies (5.5). Clearly, we can assume $r<\delta$, i.e., $[\delta, T-\delta]\subset[r, T-r]$ and thus it is sufficient to show that

$$\|(AF^{(\varepsilon)}) * \rho_r - A(F^{(\varepsilon)} * \rho_r)\|_\gamma \to 0 \quad \text{and} \quad \|f * \rho_r - f\|_\gamma \to 0.$$

We have

$$\begin{aligned} &\frac{1}{\gamma(t, x)}[(AF^{(\varepsilon)} * \rho_r - A(F^{(\varepsilon)} * \rho_r)](t, x, y) \\ &= \frac{1}{\gamma(t, x)} \iint \Big\{ \sum_{i,j=1}^n [a_{ij}(t+\sigma, x+\xi, y) - a_{ij}(t, x, y)] F^{(\varepsilon)}_{x_i x_j}(t+\sigma, x+\xi) \\ &\qquad + \sum_{i=1}^n [b_i(t+\sigma, x+\xi, y) - b_i(t, x, y)] F^{(\varepsilon)}_{x_i}(t+\sigma, x+\xi) \\ &\qquad - [c(t+\sigma, x+\xi, y) - c(t, x, y)] F^{(\varepsilon)}(t+\sigma, x+\xi) \Big\} \rho_r(\sigma, \xi)\, d\sigma\, d\xi \\ &\leqq \sum \tilde{a}_{ij}(r) \|F^{(\varepsilon)}_{x_i x_j}\|_\gamma + \sum \tilde{b}_{ij}(r) \|F^{(\varepsilon)}_{x_i}\|_\gamma + \tilde{c}(r) \|F^{(\varepsilon)}\|_\gamma \end{aligned} \tag{5.8}$$

where $\tilde{a}_{ij}$, $\tilde{b}_i$, $\tilde{c}$ denote the moduli of continuity of the corresponding coefficients. Since the coefficients were assumed to be uniformly continuous and the $\|\cdot\|_\gamma$-norms of $F^{(\varepsilon)}_{x_i x_j}$,

$F_{x_i}^{(\varepsilon)}$, and $F^{(\varepsilon)}$ are finite by the assumption of the theorem, the right-hand side of the inequality tends to zero as $r \to 0$, thus proving the lemma.

LEMMA 5.2.

$$\int_{[t_1,t_2]\times \mathbf{R}^n \times Y} \gamma \, dM \leqq (t_2 - t_1) \cdot \gamma(t, x) \tag{5.9}$$

holds true for every $M \in \mathfrak{M}^W(t, x)$ *and* $0 \leqq t_1 \leqq t_2 \leqq T$.

Proof. Denote $\chi(s) := (t_2 - t_1) - \int_0^s 1_{[t_1,t_2]}(\sigma)\, d\sigma$ and let $\chi_k : [0, T] \to \mathbf{R}^1$ be a monotonely decreasing sequence of functions that is continuously differentiable in $(0, T)$, for which $\chi_k(T) = 0$ and such that $\chi_k \searrow \chi$ and $\chi_k' \nearrow -1_{[t_1, t_2]} = \chi'$ as $k \to \infty$.

For $M \in \mathfrak{M}^W(t, x) \subset \mathcal{M}_A(t, x)$, the generalized Dynkin formula (3.2) can be applied to the functions $\Phi_k(\sigma, \xi) := \chi_k(\sigma) \cdot \gamma(\sigma, \xi)$. Using relation

$$A\Phi_k = \gamma \cdot A\chi_k + \chi_k \cdot A\gamma = \gamma \cdot \chi_k' - \gamma \cdot c \cdot \chi_k + \chi_k \cdot A\gamma$$

and the fact that $c \cdot \gamma$, $-A\gamma$, and M are nonnegative, we obtain

$$\begin{aligned}(t_2 - t_1) \cdot \gamma(t, x) &\geqq \gamma(t, x) \cdot \chi(t) = \lim_{k\to\infty} \gamma(t, x) \cdot \chi_k(t) \\ &= \lim_{k\to\infty} \Phi_k(t, x) = \lim_{k\to\infty} \int -A\Phi_k \, dM \\ &= \lim_{k\to\infty} \int -\chi_k' \cdot \gamma \, dM + \int \chi_k c \cdot \gamma \, dM + \int \chi_k(-A\gamma)\, dM \\ &\geqq \lim_{k\to\infty} \int -\chi_k' \cdot \gamma \, dM = \int 1_{[t_1,t_2]} \cdot \gamma \, dM,\end{aligned}$$

thus proving the lemma.

Conclusion of the Proof of Theorem 2. Since $F^{(\varepsilon,\delta)} \in C_\gamma^2(E)$ and $M_0 \in \mathfrak{M}^W \subset \mathcal{M}_A(t, x)$, we can apply the generalized Dynkin formula and by (5.5) and (5.9) obtain

$$\begin{aligned}F^{(\varepsilon,\delta)}(t, x) &= -\int_{[0,T]\times \mathcal{R}^n \times Y} AF^{(\varepsilon,\delta)} \, dM_0 \\ &\leqq \int_{[\delta, T-\delta]\times \mathbf{R}^n \times Y} f \, dM_0 + \delta \cdot \int_{[\delta, T-\delta]\times \mathbf{R}^n \times Y} \gamma \, dM_0 \\ &\quad + \|AF^{(\varepsilon,\delta)}\|_\gamma \cdot \int_{([0,\delta]\cup[T-\delta,T])\times \mathbf{R}^n \times Y} \gamma \, dM_0.\end{aligned}$$

Since $M_0 \in \mathfrak{M}^W \subset \mathcal{M}^{\gamma,\Gamma}$, we have $0 \leqq \int_S \gamma \, dM_0 \leqq \Gamma$. From Lemma 5.2 it follows that the integral in the last term is not greater than $2\delta \cdot \gamma(t, x) = 2\delta \cdot \alpha \cdot \Gamma$ and since by Lemma 5.1 $\|AF^{(\varepsilon,\delta)}\|_\gamma \leqq \|AF^{(\varepsilon)}\|_\gamma + \delta$, we have

$$F^{(\varepsilon,\delta)}(t, x) \leqq \int_S f \, dM_0 + \delta \cdot 2(1 + \alpha \|AF^{(\varepsilon)}\|_\gamma) \cdot \Gamma. \tag{5.10}$$

Choosing first ε then δ sufficiently small, from $\|F - F^{\varepsilon,\delta}\| \leqq \varepsilon + \delta$ and relation (5.10) it follows that

$$F(t, x) = \inf\left\{\int f \, dM^u : M^u \in \mathfrak{M}^S\right\} \leqq \int f \, dM_0$$

in contradiction to the choice (5.2) of f as a separating functional. This proves the equivalence of the strong and weak formulations.

Remark. Assumptions on the derivatives of $F^{(\varepsilon)}$ were only needed to obtain estimate (5.8). Note that since $F^{(\varepsilon)}$ is locally Lipschitzian, its first derivatives exist

almost everywhere and are locally bounded. This fact alone is sufficient to prove the equivalence of the Strong and Weak Problems provided the diffusion coefficients a_{ij} do not depend on t and x. In fact, in this case the terms $[a_{ij}(t+\sigma, x+\xi, y)-a_{ij}(t, x, y)]$ are zero and no assumptions on the second derivatives $F_{x_i x_j}$ are needed, shortcutting the approximation by $F^{(\varepsilon)}$ and the entire § 6.

COROLLARY 1. *Suppose that l is of at most linear growth, i.e., $|l(t, x, y)| \leqq r_0+r_1|x|+r_2 t$. If the processes are deterministic ($a_{ij} \equiv 0$) or the diffusion coefficients are independent of time and space, then the strong and weak problems are equivalent.*

Proof. If l is of linear growth, then γ can be chosen to be $\bar{r} \cdot (1+|x|+t)$ with $\bar{r} > \max(r_0, r_1, r_2)$. Consequently, f will be uniformly Lipschitzian and so will F. Moreover, F can be represented as the sum of a concave and of a smooth function, and hence the first and second partial derivatives of F exist almost everywhere and the first partials are uniformly bounded. The corollary then follows from the previous remark.

The measure M_0 introduced in the proof of Theorem 2 could not be in the w^* convex closure of $\mathfrak{M}^S(t_0, x_0)$. The argument there in fact proves the following corollary.

COROLLARY 2. *$\mathfrak{M}^W(t, x)$ is the w^* convex closure of $\mathfrak{M}^S(t, x)$.*

6. A Sobolev approximation of the value function. To complete the proof of the equivalence of the Strong and Weak Problems, it remains to show that the value function generated by a smooth running cost can be approximated by a $W_\infty^{1,2}$ function in the way required by the assumptions of Theorem 2. This kind of approximability of the value function that does not use any nondegeneracy assumptions is also of independent interest in other branches of control theory unrelated to the strong and weak formulations. This section is devoted to the proof of the result.

THEOREM 3. *Let $f \in C_\gamma^2(S)$ and let F be the corresponding value function defined by* (5.3). *Then for every $\varepsilon > 0$ there exists a function $F^{(\varepsilon)} \in C_\gamma(E)$ with the following properties*:

(a) $\|F-F^{(\varepsilon)}\|_\gamma \leqq \varepsilon$;

(b) *The partial derivatives $F_t^{(\varepsilon)}$, $F_{x_i}^{(\varepsilon)}$, $F_{x_i, x_j}^{(\varepsilon)}$ exist almost everywhere for every $1 \leqq i, j \leqq n$ and satisfying $|||F^{(\varepsilon)}|||_\gamma \leqq K(\varepsilon)$ with some constant $K(\varepsilon)$ where*

$$|||F^{(\varepsilon)}|||_\gamma := \|F^{(\varepsilon)}\|_\gamma + \|F_t^{(\varepsilon)}\|_\gamma + \sum_{i=1}^n \|F_{x_i}^{(\varepsilon)}\|_\gamma + \sum_{i=1}^n \|F_{x_i x_j}^{(\varepsilon)}\|_\gamma;$$

(c) $A^y F^{(\varepsilon)}(t, x)+f(t, x) \geqq 0$ *for almost every* $(t, x) \in E$, *for every* y, *and* $F^{(\varepsilon)}(T, x) = 0$ *whenever* $T < \infty$.

We denote the weighted Sobolev space of all functions satisfying (b) by $W_{\gamma,\infty}^{1,2}$.

The idea of the proof is to extend the control set of the original problem by one additional "smoothing control" giving rise to an n-dimensional Brownian motion. The value function of the extended problem will then have the required smoothness properties and, by charging a sufficiently high penalty for the "smoothing," its domain of application can be kept small and this way the smoothed value function can be forced to remain close to the original one.

To be more precise, let us introduce one more additional control η so that the extended control set will be $Y \cup \{\eta\}$. The process associated with η will be the standard n-dimensional Brownian motion discounted at the lowest possible rate $c_0 = \inf_{t,x,y} c(t, x, y)$ so that we have

$$E_{t,x}^\eta \Phi(x_{t+s}^\eta) = \frac{e^{-c_0 s}}{(2\pi s)^{n/2}} \int \Phi(t+s, \xi) \exp\left[-\frac{|\xi-x|^2}{2s}\right] d\xi =: (\beta_s * \Phi)(t, x).$$

The infinitesimal operator corresponding to the exponentially killed Brownian motion is

$$A^{\eta}\Phi = \tfrac{1}{2}\Delta\Phi - c_0\Phi$$

where Δ denotes the Laplacian. Recall that in Lemma 2.1 inequality (2.2) holds not only for the family of operators $\{A^{\gamma}\}_{\gamma\in Y}$ but with possibly different numbers α and $\bar{\alpha}$ also for the extended family $\{A^{\gamma}\}_{\gamma\in Y\cup\{\eta\}}$. In particular, we have $0<\alpha\gamma\leqq -A^{\eta}\gamma$.

During the period of time when the new control η is applied we charge the running cost

$$f(s,x,\eta) := L\cdot(-A^{\eta}\gamma)(s,x) = L\cdot(c_0\gamma(s,x)-\tfrac{1}{2}\Delta\gamma(s,x))$$

with some constant L to be determined later. For simplicity we only allow η to be applied during at most one nonrandom interval of time. In other words, the extended set $\mathcal{U}_{\eta}$ of admissible controls will be the set of all functions of the form

$$v(\omega,s)=\begin{cases}\eta & \text{if } t_1\leqq s<t_2,\\ u(\omega,s) & \text{otherwise,}\end{cases}$$

with all possible choices of $0\leqq t_1\leqq t_2\leqq T$ and $u\in\mathcal{U}$. Note that because of the possibility of killing, the processes may die before t_1 or t_2.

The value function of the extended problem can then be written as

$$\begin{aligned}F^{L}(t,x) &:= \inf_{v\in\mathcal{U}_{\eta}} E^{v}_{t,x}\int_t^T [f(s,x^{v}_s,v_s)1_{Y}(v_s)+L\cdot(-A^{\eta}\gamma)(s,x_s)1_{\{\eta\}}(v_s)]\,ds\\ &= \inf_{\substack{t\leqq t_1\leqq t_2\leqq T\\ u\in\mathcal{U}}} E^{u}_{t,x}\Bigg\{\int_t^{t_1} f(s,x^{u}_s,u_s)\,ds\\ &\qquad + E^{\eta}_{t_1,x^{u}_{t_1}}\Bigg[\int_{t_1}^{t_2} L\cdot(-A^{\eta}\gamma)(s,x^{\eta}_s)\,ds+F(t_2,x^{\eta}_{t_2})\Bigg]\Bigg\} \qquad (6.1)\\ &\leqq \inf_{\substack{t\leqq t_1\leqq t_2\leqq T\\ u\in\mathcal{U}}} E^{u}_{t,x}\Bigg\{\int_t^{t_1} f(s,x^{u}_s,u_s)\,ds+(\beta_{t_2-t_1}*F)(t_1,x^{u}_{t_1})\\ &\qquad + L\cdot(\gamma-\beta_{t_2-t_1}*\gamma)(t_1,x^{u}_{t_1})\Bigg\}.\end{aligned}$$

Now we show that setting the penalty L high will keep the optimal cost F^{L} close to F.

PROPOSITION 6.1. *For every $\varepsilon>0$ there exists a $0\leqq L_{\varepsilon}<\infty$ such that $\|F-F^{L_{\varepsilon}}\|_{\gamma}\leqq\varepsilon$.*

Proof. $F^{L}\leqq F$ is trivial since Φ^{L} is the value functional of the extended control problem that contains the original problem embedded, as $t_2=t_1$ is permitted.

To show that $F-F^{L_{\varepsilon}}\leqq\varepsilon\cdot\gamma$, observe that since F/γ is bounded and uniformly continuous there exists a t_{ε} such that $\|F/\gamma-\beta_h*(F/\gamma)\|\leqq\varepsilon/2$ for all $0\leqq h<t_{\varepsilon}$. With this t_{ε} let us choose $L_{\varepsilon}:=3\|f\|_{\gamma}/(\alpha\cdot t_{\varepsilon})$. Now let us consider an arbitrary $(t,x)\in E$. Since $F^{L_{\varepsilon}}$ is the pointwise infimum in (6.1), we can find an $\varepsilon/2$-optimal triple $\bar{u}$, $0\leqq\bar{t}_1\leqq\bar{t}_2$, i.e., such that

$$\begin{aligned}\bar{F}(t,x) &:= E^{\bar{u}}_{t,x}\Bigg\{\int_t^{\tau_1} f(\bar{x}_s,\bar{u}_s)+(\beta_h*F)(\bar{t}_1,\bar{x}_{\bar{t}_1})+L_{\varepsilon}\cdot E^{\beta}_{\bar{t}_1,\bar{x}_{\bar{t}_1}}\int_{\bar{t}_1}^{\bar{t}_2}(-A^{\eta}\gamma)(w_s)\,ds\Bigg\}\\ &\leqq F^{L_{\varepsilon}}(t,x)+\frac{\varepsilon}{2}\cdot\gamma(t,x).\end{aligned}$$

We use the notation $\bar{x}_s = x_s^{\bar{u}}$ and $h = \bar{t}_2 - \bar{t}_1$. Keep in mind that although $\bar{u}$, $\bar{t}_1$, $\bar{t}_2$ do depend on (t, x), the numbers t_ε and L_ε were chosen before (t, x) was picked, hence estimates involving only t_ε and L_ε will hold for every $(t, x) \in E$.

With the quantities just defined we can write

$$\begin{aligned} F(t,x) - F^{L_\varepsilon}(t,x) &= [F(t,x) - \bar{F}(t,x)] + [\bar{F}(t,x) - F^{L_\varepsilon}(t,x)] \\ &\leqq [F(t,x) - \bar{F}(t,x)] + \gamma(t,x) \cdot \varepsilon/2 \end{aligned} \tag{6.2}$$

and it remains to show that $F(t, x) - \bar{F}(t, x) \leqq \gamma(t, x) \cdot \varepsilon/2$.

We increase the value if we fix $\bar{u}$ for the initial interval $[t, \bar{t}_1)$ and allow minimization only after $\bar{t}_1$:

$$\begin{aligned} F(t,x) - \bar{F}(t,x) &\leqq E_{t,x}^{\bar{u}} \left\{ \int_t^{\bar{t}_1} f(s, \bar{x}_s, \bar{u}_s)\, ds + F(\bar{t}_1, \bar{x}_{\bar{t}_1}) \right\} - \bar{F}(t,x) \\ &= E_{t,x}^{\bar{u}} \{ ([F - \beta_h * F] - L_\varepsilon \cdot [\gamma - \beta_h * \gamma])(\bar{t}_1, \bar{x}_{\bar{t}_1}) \}. \end{aligned} \tag{6.3}$$

The expression in the first bracket under the expectation sign normalized by γ can be estimated at an arbitrary $(t^1, x^1) \in E$ as

$$\begin{aligned} \frac{1}{\gamma(t', x')} [F(t', x') - (\beta_h * F)(t', x')] &= \left(\frac{F}{\gamma} - \beta * \frac{F}{\gamma} \right)(t', x') + \left(\beta * \frac{F}{\gamma} - \frac{\beta * F}{\gamma} \right)(t', x') \\ &\leqq \left\| \frac{F}{\gamma} - \beta * \left(\frac{F}{\gamma} \right) \right\| + \|F\|_\gamma \cdot \left(1 - \frac{\beta_h * \gamma(t', x')}{\gamma(t', x')} \right). \end{aligned}$$

Consequently, if we divide the whole expression under the expectation in (6.3) by γ we obtain for it

$$\begin{aligned} \chi(t', x') &:= \frac{1}{\gamma(t', x')} \cdot ([F - \beta_h * F] - L_\varepsilon \cdot [\gamma - \beta_h * \gamma])(t', x') \\ &\leqq \left\| \frac{F}{\gamma} - \beta_h * \frac{F}{\gamma} \right\| - (L_\varepsilon - \|F\|_\gamma) \cdot \left(1 - \frac{\beta_h * \gamma(t', x')}{\gamma(t', x')} \right). \end{aligned} \tag{6.4}$$

Now there are two possibilities: either $h \leqq t_\varepsilon$ or $h > t_\varepsilon$. If $h \leqq t_\varepsilon$ then by the definition of t_ε we have $\|F/\gamma - \beta_h * (F/\gamma)\| \leqq \varepsilon/2$. Since $L_\varepsilon \geqq \|f\|_\gamma/\alpha \geqq \|F\|_\gamma$ and $\gamma \geqq \beta_h * \gamma$, the last term is nonnegative; we may subtract it and we get $\chi(t', x') \leqq \varepsilon/2$ for arbitrary $(t', x') \in E$.

On the other hand, if $h > t_\varepsilon$, then $1 - e^{-\alpha h} \geqq 1 - e^{\alpha t_\varepsilon}$; thus by Corollary 3 to Lemma 2.1 and the choice of L_ε, we have

$$(L_\varepsilon - \|F\|_\gamma) \cdot \left(1 - \frac{\beta_h * \gamma(t', x')}{\gamma(t', x')} \right) \geqq \frac{2\|f\|_\gamma}{\alpha(1 - e^{-\alpha t_\varepsilon})} \cdot (1 - e^{-\alpha h}) \geqq \frac{2\|f\|_\gamma}{\alpha} \geqq 2\|F\|_\gamma.$$

Since both F/γ and $\beta_h * (F/\gamma)$ are bounded by $\|F\|_\gamma$, from (6.4) we find that $\chi(t', x') \leqq 0 \leqq \varepsilon/2$ for every $(t', x') \in E$.

Substituting this result back in (6.3) we find by Corollary 3 to Lemma 2.1 that

$$F(t,x) - \bar{F}(t,x) \leqq E_{t,x}^{\bar{u}} \gamma(t_1, \bar{x}_{t_1}) \cdot \chi(t_1, \bar{x}_{t_1}) \leqq \frac{\varepsilon}{2} E_{t,x}^{\bar{u}} \gamma(t_1, \bar{x}_{t_1}) \leqq \frac{\varepsilon}{2} \cdot \gamma(t,x).$$

This, together with (6.2) gives $F - F^{L_\varepsilon} \leqq \varepsilon \cdot \gamma$, which completes the proof of the proposition.

It is well known (cf., e.g., [2, Thm. 4.2]) that under the conditions of Theorem 3 the value function F^L permits the decomposition $F^L = \tilde{F}^L + \hat{F}^L$ where $\tilde{F}^L \in C_\gamma$ is smooth, its partial derivatives $\tilde{F}^L_t$, $\tilde{F}^L_{x_i}$, $\tilde{F}^L_{x_i x_j}$ belong to C_γ while $\hat{F}^L \in C_\gamma$ is concave in x and monotone in t. In fact, for every control $J^u \in C^2_\gamma$, the infimum of continuously parametrized family of C^2_γ functions has the above decomposition property. For such functions the generalization of Alexandrov's Theorem [1] holds true; for almost every (t, x) the derivatives F^L_t, $F^L_{x_i}$, $F^L_{x_i x_j}$ exist and satisfy

$$\begin{aligned} F^L(t+\sigma, x+\xi) = F^L(t, x) + F^L_t(t, x)\cdot\sigma + \sum F^L_{x_i}(t, x)\xi_i \\ + \sum\sum F^L_{x_i x_j}(t, x)\xi_i\xi_j + o(|t|+|\xi|^2). \end{aligned} \tag{6.5}$$

It is easy to see that F^L satisfies the Hamilton-Jacobi-Bellman inequality of the extended problem almost everywhere. In fact, the next proposition is only a slight modification of known results (cf. [2], [3]) that we prove here only because the easy proof makes our exposition self-contained.

PROPOSITION 6.2. *For every* $y \in Y \cup \{\eta\}$,

$$A^y F^L(t, x) + f(t, x, y)\cdot 1_Y(y) + L\cdot(-A^y\gamma)(t, x)1_{\{\eta\}}(y) \geqq 0 \tag{6.6}$$

for almost every $(t, x) \in E$.

Proof. Suppose there exists a $y \in Y_0$ and a $(t_0, x_0) \in E$ from the nonexceptional set such that

$$A^{Y_0} F^L(t_0, x_0) + f(t_0, x_0, y_0) \leqq -\delta < 0.$$

Then, by the continuity of the underlying processes, there exists an $s_0 > 0$ such that for all $s \leqq s_0$

$$s^{-1}[E^{y_0}_{t_0, x_0} F^L(t_0 + s, x_s) - F^L(t_0, x_0)] + f(t_0, x_0, y_0) \leqq -\delta/2 < 0.$$

Let $u^\delta_\xi \in \mathcal{U}_\eta$ be a $\delta_{s_0}/3$-optimal control for the initial state $(t_0 + s_0, \xi)$ and define

$$u^0(\omega, r) := \begin{cases} y_0 & \text{if } t_0 \leqq r < t_0 + s_0, \\ u^\delta_\xi & \text{if } r \geqq t_0 + s_0 \text{ and } x_{t_0+s_0}(\omega) = \xi. \end{cases}$$

Then this control is again in $\mathcal{U}_\eta$ and will yield the cost

$$\begin{aligned} E^{u_0}_{t_0, x_0} \int_{t_0}^{\tau} f(s, x_s, u_0(s))\, ds &\leqq f(t_0, x_0, y_0)\cdot s_0 + E^{y_0}_{t_0, x_0}\{F^L(t_0 + s, x_{t_0+s_0}) + \delta_{s_0}/3\} \\ &\leqq F^L(t_0, x_0) - \delta_{s_0}/6 < 0 \end{aligned}$$

in contradiction to the definition of F^L as the infimum over all $u \in \mathcal{U}_\eta$. The proof for $y_0 = \eta$ is the same.

Conclusion of the Proof of Theorem 3. Let us choose $F^{(\varepsilon)} := F^{L_\varepsilon}$ according to Proposition 6.1. Then we have $\|F - F^{(\varepsilon)}\|_\gamma \leqq \varepsilon$. The derivatives $F^{(\varepsilon)}_t$, $F^{(\varepsilon)}_{x_i}$, $F^{(\varepsilon)}_{x_i, x_j}$ exist almost everywhere by Alexandrov's Theorem, and Proposition 6.2 shows that the Hamilton-Jacobi inequality holds true for every $y \in Y$ and for almost every $(t, x) \in E$. The smooth component $\tilde{F}^{(\varepsilon)}$ and its derivatives are in C_γ by Krylov's cited result [2, Thm. 4.2].

It remains to show that the derivatives of the concave component $\hat{F}^{(\varepsilon)}$ are essentially bounded by $K(\varepsilon)\cdot\gamma$.

Consider the first derivative in an arbitrary direction of the (t, x)-space. By the concavity of $\hat{F}^{(\varepsilon)}$ this directional derivative is monotone along each line parallel to the chosen direction. Suppose that this (one-dimensional) derivative function exceeds $K\cdot\gamma$ for every K. Then by its monotonicity it follows that neither can its integral

function be bounded by $K_1 \cdot \gamma$. But this contradicts $\hat{F}^{(\varepsilon)} \in C_\gamma$, which follows from Proposition 6.1. Hence there must be a $K_2(\varepsilon)$ such that $|\hat{F}_t^{(\varepsilon)}(t,x)| + \sum_{i=1}^n |\hat{F}_{x_i}^{(\varepsilon)}(t,x)| \leqq K_2(\varepsilon)\gamma(t,x)$ almost everywhere.

As for the second derivatives, $\hat{F}_{x_i,x_j}^{(\varepsilon)}(t,x) \leqq K_3(\varepsilon)$ follows from the concavity of $\hat{F}^{(\varepsilon)}$. To show that $|F_{x_i x_j}^{(\varepsilon)}| \leqq K(\varepsilon) \cdot \gamma$, consider inequality (6.6) of Proposition 6.2 for $y = \eta$. This claims that

$$F_t^{(\varepsilon)} + \tfrac{1}{2}\Delta F^{(\varepsilon)} - c_0 F^{(\varepsilon)} \geqq L_\varepsilon \cdot A^\eta \gamma \geqq L_\varepsilon \cdot \bar{\alpha} \cdot \gamma$$

where the last inequality follows from the right-hand side of (2.2). Using the estimates already obtained for $F^{(\varepsilon)}$, $F_t^{(\varepsilon)}$, $\tilde{F}_{x_i,x_j}^{(\varepsilon)}$ we get

$$\begin{aligned}\Delta \hat{F}^{(\varepsilon)}(t,x) &\geqq -2\left(c_0 \|F^{(\varepsilon)}\|_\gamma + \|F_t^{(\varepsilon)}\|_\gamma + \frac{1}{2}\sum_{ij=1}^n \|\tilde{F}_{x_i x_j}\|_\gamma + L_\varepsilon \bar{\alpha} \right) \cdot \gamma(t,x) \\ &= -K_4(\varepsilon)\gamma(t,x).\end{aligned}$$

This lower bound for the sum $\sum \hat{F}_{x_i x_j}^{(\varepsilon)}/\gamma$ together with the upper bound for the individual summands $\hat{F}_{x_i,x_j}^{(\varepsilon)}$ obtained from the concavity of $\hat{F}^{(\varepsilon)}$ implies that $\|F_{x_i,x_j}^{(\varepsilon)}\|_\gamma \leqq K(\varepsilon)$ for every $1 \leqq i, j \leqq n$. This completes the proof of Theorem 3.

7. Semicontinuous costs. In the previous paragraphs, in particular in §§ 3 and 4, we assumed the running cost to be continuous $l \in C_\gamma$. Now we are going to remove this assumption and allow l to be lower semicontinuous and of growth less than γ. More precisely we denote by LC_γ the set of all functions l satisfying the following:

(i) l is lower semicontinuous;

(ii) $\sup_{(\theta,y)\in\mathcal{S}} |l(\xi,y)|/\gamma(\xi) < \infty$;

(iii) $\limsup_{|\xi|\to\infty} l(\xi,y)/\gamma(\xi) = 0$.

Such functions can be represented as upper envelopes of continuous functions $l = \sup\{f : f \in C_\gamma, f \leqq l\}$ or even as limits of nondecreasing sequences of C_γ functions.

The aim of the present paragraph is to show that all results proved for continuous l in the preceding paragraphs remain true for control problems with lower semicontinuous cost functions $l \in LC_\gamma$. The key tool in approximating lower semicontinuous costs by continuous ones will be the following min-max type argument.

PROPOSITION 7.1. *Suppose $l \in LC_\gamma$ and let $\mathcal{K}$ denote an arbitrary w^*-compact subset of $\mathcal{M}_+^\gamma(\mathcal{S})$. Then*

$$\inf_{\mu\in\mathcal{K}} \int l\,d\mu = \inf_{\mu\in\mathcal{K}} \sup_{f\leqq l, f\in C_\gamma} \int f\,d\mu = \sup_{f\leqq l, f\in C_\gamma} \inf_{\mu\in\mathcal{K}} \int f\,d\mu. \tag{7.1}$$

Proof. Note first, that every $l \in LC_\gamma$ defines a convex, lower w^*-semicontinuous functional on $\mathcal{M}_+^\gamma$, and hence all infima in (7.1) are attained for some elements of the w^*-compact set $\mathcal{K}$.

The Monotone Convergence Theorem and the obvious inequality $\inf\sup \geqq \sup\inf$ yield

$$I_0 := \inf_{\mu\in\mathcal{K}} \int l\,d\mu = \inf_{\mu\in\mathcal{K}} \sup_{f\leqq l, f\in C_\gamma} \int f\,d\mu \geqq \sup_{f\leqq l, f\in C_\gamma} \inf_{\mu\in\mathcal{K}} \int f\,d\mu.$$

Let μ^f denote the measure, for which $\int f\,d\mu^f = \inf_{\mu\in\mathcal{K}} \int f\,d\mu$. To prove the proposition it is sufficient to show the existence of a $\mu^* \in \mathcal{K}$ for which

$$\sup_{f\leqq l, f\in C_\gamma} \int f\,d\mu^f \geqq \int l\,d\mu^* \tag{7.2}$$

holds true.

Let f_k denote a monotone nondecreasing sequence of continuous functions with $f_k \in C_\gamma(S)$ and $f_k \nearrow l$ as $k \to \infty$. Since $\mathcal{K}$ is sequentially compact, we can select a subsequence k_i such that $\mu_i := \mu^{f_{k_i}}$ converge weakly* to a limit $\mu^* \in \mathcal{K}$ as $i \to \infty$.

Let us consider the following array of reals:

$$I(i, j) := \int f_{k_i}\, d\mu_j, \qquad i, j = 1, 2, \cdots.$$

If $i' < i$ then $I(i', j) \leqq I(i, j)$ because the sequence f_{k_i} is monotone nondecreasing. The measure μ_i is by definition minimizing $\int f_{k_i}\, d\mu$ and as $f_{k_i} \leqq l$, we have

$$I(i, i) = \int f_{k_i}\, d\mu_i = \inf_{\mu \in \mathcal{K}} \int f_{k_i}\, d\mu \leqq \inf_{\mu \in \mathcal{K}} \int l\, d\mu = I_0.$$

Consequently all elements $I(i, j)$ with $i \leqq j$ (i.e., above the diagonal) are uniformly bounded by I_0. From the monotonicity of the sequence f_{k_i} it follows that the diagonal sequence $I(i, i)$ is monotone nondecreasing and so $I_\infty := \lim_{i \to \infty} I(i, i) \leqq I_0$ exists.

Since f_{k_i} is continuous and $\mu^* = w^* - \lim \mu_j$, it follows that the sequence $I(i, j)$ converges for any fixed i to a limit $I(i, \infty) = \int f_{k_i}\, d\mu^*$ as $j \to \infty$. From $I(i, j) \leqq I(j, j)$ for $i \leqq j$ it follows that

$$I(i, \infty) = \lim_{j \to \infty} I(i, j) \leqq \lim_{j \to \infty} I(j, j) = I_\infty.$$

Recall that the sequence f_k was chosen such that $f_k \nearrow l$. Consequently the monotone convergence theorem yields

$$I(i, \infty) = \int f_{k_i}\, d\mu^* \nearrow \int l\, d\mu^* \leqq I_\infty = \sup_{f \leqq l, f \in C_\gamma} \int f\, d\mu^f.$$

In other words μ^* satisfies (7.2) and the proof is complete.

THEOREM 4. *Suppose $l \in LC_\gamma$. Then the (strong) value function of the stochastic control problem formulated in §§ 1-2 is the upper envelope of the smooth subsolutions of the Hamilton-Jacobi-Bellman equation, i.e.,*

$$\Psi(t, x) = \sup \{\Phi(t, x) \colon \Phi \in C_\gamma^2,\ A\Phi + l \geqq 0\}. \tag{7.3}$$

Proof. It was shown in §§ 5-6 that $\mathfrak{M}^W$ is the closed convex hull of $\mathfrak{M}^S$. Since $\int l\, dM$ is a convex, lower w^*-semicontinuous functional on $\mathcal{M}_\pm^\gamma$ whenever $l \in LC_\gamma$, it follows that its infimum over $\mathfrak{M}^S$ is the same as its minimum attained in $\mathfrak{M}^W$. Consequently the strong and weak value functions coincide even if l is only lower semicontinuous $l \in LC_\gamma$.

We know from Theorem 1 that the value function permits representation (7.3) if l is continuous ($l \in C_\gamma$). Proposition 7.1 can be applied to $l \in LC_\gamma$ and $\mathcal{K} = \mathfrak{M}^W$ as $\mathfrak{M}^W$ is a w^*-compact set and we obtain

$$\begin{aligned}\Psi(t, x) &= \inf_{M \in \mathfrak{M}^W(t,x)} \int l\, dM = \sup_{f \leqq l, f \in C_\gamma} \inf_{M \in \mathfrak{M}^W} \int f\, dM \\ &= \sup_{f \leqq l, f \in C_\gamma} \sup \{\Phi(t, x) \colon \Phi \in C_\gamma^2(E),\ A\Phi + f \geqq 0\} \\ &\leqq \sup \{\Phi(t, x) \colon \Phi \in C_\gamma^2,\ A\Phi + l \geqq 0\}.\end{aligned}$$

The opposite inequality is immediate, since for every $\Phi \in C_\gamma^2$ with $A\Phi + l \geqq 0$ and for every $M \in \mathfrak{M}^W = \mathcal{M}_\gamma^\Gamma \cap \mathcal{M}_A$ Dynkin's formula yields

$$\Phi(t, x) = \int (-A\Phi)\, dM \leqq \int l\, dM.$$

Taking infimum over $M \in \mathfrak{M}^W$, we obtain $\Phi(t, x) \leqq \inf_{M \in \mathfrak{M}^W} \int l\, dM = \psi(t, x)$ for every $\Phi \in C_\gamma^2$ with $A\Phi + l \geqq 0$, which completes the proof of the theorem.

8. Inclusion of terminal penalties. In this final paragraph we explain how to extend the main results of the paper to problems where the cost function also includes an additional terminal penalty, i.e., where the objective is to minimize the functional

$$J^u(t, x) = E_{t,x}^u \left\{ \int_t^T l(t, x_t, m_t)\, dt + L(x_T) \right\} \qquad (T < \infty) \tag{8.1}$$

over all controls $u \in \mathcal{U}$. Here both l and L are lower semicontinuous functions of growth less than γ at infinity. This will extend the scope of the results to include problems like the maximization of the hitting probability of a closed target set or the fixed endpoint problem of deterministic control theory that were beyond the reaches of the other approaches to the Hamilton–Jacobi theory.

The key to the extension is to consider a more elaborate state space $\tilde{S}$ that is composed of S_0, the compactification of the "interior" of the state-space, and of S_∂, the compactified "terminal boundary," as two separate components. More precisely, let $\tilde{S}$ denote the compact metric space that consists of the two isolated subsets $S_0 := E \times Y$ and $S_\partial := \bar{\mathbf{R}}^n$. Note that S_0 is the same space that was denoted by S in § 2.

Every continuous function $\Phi \in C_\gamma(\tilde{S})$ will then correspond to the pair $\Phi|_{S_0} \in C_\gamma(S_0)$ and $\Phi|_{S_\partial} \in C_\gamma(S_\partial)$ where $C_\gamma(S_0)$ is $C_\gamma(S)$ of § 2 and

$$C_\gamma(S_\partial) := \left\{ \phi \in C(\mathbf{R}^n) \colon \sup_{x \in \mathbf{R}^n} |\phi(x)|/\gamma(T, x) < \infty \quad \text{and} \quad \lim_{|x| \to \infty} |\phi(x)|/\gamma(T, x) = 0 \right\}.$$

$LC_\gamma(\tilde{S})$ will denote the set of all lower semicontinuous functions on $\tilde{S}$, i.e., those that can be represented as upper envelopes of families of $C_\gamma(\tilde{S})$ functions. The dual space to $C_\gamma(\tilde{S})$ will be the set $\mathcal{M}_\pm^\gamma(\tilde{S})$ of all pairs of measures $M = (M_0, M_\partial)$ with $M_0 \in \mathcal{M}_\pm^\gamma(S_0)$, $M_\partial \in \mathcal{M}_\pm^\gamma(S_\partial)$ provided with the norm $\|M\|_\gamma = \int \gamma(t', x') |M_0(dt', dx', dy)| + \int \gamma(T, x) |M_\partial(dx)|$. The set of all nonnegative measures $M \in \mathcal{M}_\pm^\gamma(\tilde{S})$ with $\|M\|_\gamma \leqq \Gamma < +\infty$ will be denoted by $\mathcal{M}_\pm^{\gamma,\Gamma}(\tilde{S})$.

Observe that the function

$$\tilde{l}(\sigma) := \begin{cases} l(\tau, \xi, y) & \text{if } \sigma = (\tau, \xi, y) \in S_0, \\ L(x) & \text{if } \sigma = x \in S_\partial \end{cases}$$

is in $LC_\gamma(\tilde{S})$. The measure $\tilde{M}^u$ defined on the Borel sets B of $\tilde{S}$ by

$$\tilde{M}^u(B) := \begin{cases} M^u(B) & \text{of (2.1) if } B \subset S_0, \\ P_{t,x}^u(x_T^u \in B) & \text{if } B \subset S_\partial \end{cases}$$

is in $\mathcal{M}^{\gamma, 1+T}(\tilde{S}) \subset \mathcal{M}_\pm^\gamma(\tilde{S})$. With this notation the (strong) optimal control problem with both running and terminal costs can be formulated as follows:

$$\text{Minimize} \int \tilde{l}\, d\tilde{M}^u \quad \text{over all } u \in \mathcal{U}. \tag{8.2}$$

Let $C_\gamma^2(\tilde{E})$ denote the set of all twice continuously differentiable functions Φ defined on $[0, T) \times \mathbf{R}^n$ for which Φ, Φ_t, Φ_{x_i}, $\Phi_{x_i x_j}$ are all in $C_\gamma(E)$ $(i, j = 1, 2, \cdots, n)$. The difference to the definition of $C_\gamma^2(E)$ in § 2 is that now we do not require functions to vanish on the exit boundary $[T] \times \mathbf{R}^n$ for $T < \infty$. Recall that for every $\Phi \in C_\gamma^2(\tilde{E})$ Dynkin's formula

$$E_{t,x}^u \Phi(T, x_T) - \Phi(t, x) = E_{t,x}^u \int_0^T A^{u_s} \Phi(s, x_s)\, ds \tag{8.3}$$

holds true. If we introduce the operator $\tilde{A}$: $C_\gamma^2(\tilde{E}) \to C_\gamma(\tilde{S})$ by

$$\tilde{A}\Phi(\sigma) = \begin{cases} A^y\Phi(t, x) & \text{if } \sigma = (t, x, y) \in S_0, \\ -\Phi(T, x) & \text{if } \sigma = x \in S_\sigma, \end{cases}$$

then with the above notation Dynkin's formula can be written in the more compact form

$$-\Phi(t, x) = \int_{\tilde{S}} \tilde{A}\Phi \, d\tilde{M}^u. \tag{8.4}$$

The Weak Problem corresponding to (8.2) can be formulated as follows:

$$\text{Minimize} \int \tilde{l} \, d\tilde{M} \quad \text{over all } \tilde{M} \in \mathcal{M}^{\gamma, 1+T}(\tilde{S}) \cap \mathcal{M}_{\tilde{A}}(t, x)$$

with $\mathcal{M}_{\tilde{A}}(t, x) := \{\tilde{M} \in \mathcal{M}_+^\gamma(\tilde{S})$: for which (8.4) holds for all $\Phi \in C_\gamma^2(\tilde{E})\}$.

All the expositions of §§ 3–7 can be repeated word by word for this extended notation and we obtain the following theorem.

THEOREM 5. *The dual to the problem* (8.2) *is to find the supremum of all smooth subsolutions of the Hamilton–Jacobi–Bellman equation, and we have*

$$\psi(t, x) = \inf_{u \in \mathcal{U}} E_{t,x}^u \left\{ \int_t^T l(s, x_s, u_s) \, ds + L(x_T) \right\}$$

$$= \sup \Big\{ \Phi(t, x) \text{ over all } \Phi \in C_\gamma^2(\tilde{E}) \text{ satisfying}$$

$$\inf_{y \in Y} A^y\Phi(\tau, \xi) + l(\tau, \xi, y) \geqq 0 \text{ if } 0 < \tau < T, \xi \in \mathbf{R}^n$$

$$\text{and } \Phi(T, \xi) \leqq L(\xi) \quad \xi \in \mathbf{R}^n \Big\}.$$

REFERENCES

[1] A. D. ALEKSANDROV, *Almost everywhere existence of second derivatives of a convex function and related properties of convex surfaces*, Sci. Notes Leningrad State Univ., 37 (1939), pp. 3–35.

[2] N. V. KRYLOV, *Some new results from the theory of controlled diffusion processes*, Mat. Sb. (N.S.), 109 (1979), pp. 146–164.

[3] P. L. LIONS, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equation* I: *The dynamic programming principle and applications*, Comm. in Partial Differential Equations, 8 (1983), pp. 1101–1174.

[4] R. T. ROCKAFELLAR, *Extension of Fenchel's duality theorem for convex functions*, Duke Math. J., 33 (1966), pp. 81–89.

[5] D. VERMES, *Optimal control of piecewise deterministic Markov processes*, Stochastics, 14 (1985), pp. 165–208.

[6] R. B. VINTER AND R. M. LEWIS, *The equivalence of strong and weak formulations for certain problems in optimal control*, SIAM J. Control Optim., 16 (1978), pp. 546–570.

[7] ———, *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.

[8] J.-M. BISMUT, *Conjugate convex functions in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384–404.

[9] R. T. ROCKAFELLAR, *Conjugate convex functions in optimal control and the calculus of variations*, J. Math. Anal. Appl., 32 (1970), pp. 174–222.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1156-1181, September 1989

014

PIECEWISE MONOTONE FILTERING WITH SMALL OBSERVATION NOISE*

W. H. FLEMING† AND E. PARDOUX‡

This paper is dedicated to the memory of E. J. McShane.

Abstract. Nonlinear filtering of Markov diffusion processes is considered, in the case in which a piecewise monotone function of the state is observed with additive small observation noise. Under a certain detectability hypothesis, statistical tests are given to discriminate among the intervals of monotonicity during time intervals in which the state does not cross critical points of the observation function. During such time intervals, accurate approximate finite dimensional filters can be used.

Key words. nonlinear filtering, small noise, approximate finite-dimensional filters

AMS(MOS) subject classification. 93E11

1. Introduction.

There is substantial literature on the nonlinear filter model

$$\begin{aligned} dx_t &= f(x_t)\,dt + g(x_t)\,dw_t, \\ dy_t^\varepsilon &= h(x_t)\,dt + \varepsilon\,dv_t, \qquad t \geqq 0 \end{aligned} \tag{1.1}$$

where $x_t \in \mathfrak{R}^n$, $y_t \in \mathfrak{R}^l$ and w_t, v_t are independent standard Brownian motions. The random variable x_0 has distribution μ_0 and $y_0 = 0$. Finding the mean square optimal estimate $\hat{x}_t$ for x_t given $\sigma\{y_s^\varepsilon, 0 \leqq s \leqq t\}$ requires solving the nonlinear filter problem to find the conditional distribution μ_t. The dynamics of μ_t are described by the nonlinear filter equation, or by some partial differential equation equivalent to it (e.g., Zakai or pathwise filter equations [2], [7]). Thus, the nonlinear filtering problem is inherently infinite-dimensional. There are exact finite-dimensional filters only for the well-known linear case (Kalman–Bucy) and for a few special nonlinear problems (see [5] and [11]). Suppose that $\varepsilon > 0$ is small. An attractive alternative to trying to solve the nonlinear filter problem is to seek a good approximation m_t to $\hat{x}_t$, such that m_t is computable from the solution to a finite-dimensional system of stochastic differential equations driven by the observation process y_t. Moreover, we would like the number of stochastic differential equations describing the approximate filter to be small if n and l are small. There are a number of results concerning the case $n = l$ and $h(x)$ one-to-one. In this case x_t is observed exactly if $\varepsilon = 0$. For small $\varepsilon > 0$ we expect the conditional distribution μ_t to be mostly concentrated near x_t. Under some additional technical assumptions this has been shown to be correct, and good approximate filters of dimension n or $n + n^2$ have been found (see [6], [8a], [8b], and [12]–[14]). For $n = l = 1$, the simplest one-dimensional approximation m_t to $\hat{x}_t$ satisfies

$$dm_t = f(m_t)\,dt + \varepsilon^{-1}[\operatorname{sgn} h'(m_t)]g(m_t)[dy_t^\varepsilon - h(m_t)\,dt], \tag{1.2}$$

with $m_0 = Ex_0$. For $t \geqq t_0 > 0$ this gives $m_t = \hat{x}_t + O(\varepsilon)$. Certain two-dimensional approximations (m_t, R_t), where R_t is some approximation to the conditional variance, give

* Received by the editors October 7, 1988; accepted for publication (in revised form) January 10, 1989.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was partially supported by the National Science Foundation under grant MCS-8121940, by the Air Force Office of Scientific Research under contracts F-49620-86C-0111 and AFOSR-86-0315, and by the Army Research Office under contract DAAL03-86-K-0074.

‡ Mathematiques-URA225, Universite de Provence, F-13331 Marseille, Cedex 3, France and INRIA, Rocquencourt, Le Chesnay, France.

$m_t = \hat{x}_t + O(\varepsilon^{3/2})$. The accuracy is better in case $g(x)$ is constant and $h(x)$ is linear (see Picard [12, p. 1100]).

If h is not one-to-one there are substantial additional difficulties. For small $\varepsilon > 0$, $h(x_t)$ can be estimated accurately but not in general x_t itself. Moreover, when $l < n$ there are typically multiple timescales in the filtering equations (see [9] and [16]). However, when $l = n$, it often happens that enough information is contained in measurements of $h(x_t)$ plus low-intensity white noise to accurately estimate x_t, even though h is not one-to-one. It is this question with which this paper is concerned. For simplicity we consider scalar state x_t and observation $y_t (n = l = 1)$. Extensions to $n = l > 1$ are indicated in § 8. We suppose that $h(x)$ has a finite number of critical points, located at $x_1^*, \cdots, x_k^*$, with

$$h'(x_i^*) = 0, \quad h''(x_i^*) \neq 0, \quad i = 1, \cdots, k.$$

The conditional distribution μ_t may then have several peaks, located in $(-\infty, x_1^*)$, (x_k^*, ∞) and in intervals (x_i^*, x_{i+1}^*) between successive critical points. This possibility cannot always be avoided. For example, if $x_t = w_t$ and $h(-x) = h(x)$, then μ_t is symmetric about zero. However, in many instances a hypothesis test can be performed that determines, with probability very nearly one, that x_t lies in an interval of monotonicity of h. Once this is known, either an approximate filter of type (1.2) or a more accurate two-dimensional version of it can be used to estimate $\hat{x}_t$. A similar technique is used in [3] when $h(x)$ is piecewise linear and $g(x)$ is constant.

To simplify the exposition, in most of the paper the case when $h(x)$ has a single critical point ($k = 1$) is considered. Extensions of the results to $k > 1$ are discussed in § 8. Our work appears to be related to a modification of the extended Kalman filter technique to allow for Gaussian sum approximations to conditional densities that may have multiple peaks [1, § 8.4]. We shall show that under a certain "detectability" hypothesis (see (2.1), (5.10) or (8.1)) below, multiple peaks of the conditional density are improbable during time intervals in which x_t remains away from the set of critical points x_i^* of h. During such time intervals, an approximate filter of the type valid for monotone h can be used (for example, a filter of type (1.2)). Precise versions of these rough statements appear in the following sections.

The paper is organized as follows. We take the single critical point of h to be a global minimum at $x = 0$ with $h(0) = 0$. After some introductory material in §§ 2 and 3, two possible approximations m_t^+ and m_t^- to the conditional mean $\hat{x}_t$ are introduced in § 4. These satisfy equations such as (1.2) in which h is replaced by certain functions h_+ and h_-, respectively. These functions have the property that $h(x) = h_+(x)$ on a "large" interval of $\{x > 0\}$ and $h(x) = h_-(x)$ on a "large" interval of $\{x < 0\}$. Lemma 4.1 provides a test to guarantee that x_t does not cross zero during a finite time interval $a \leqq t \leqq b$, with exponentially small probability of error. Next, a test based on quadratic variations is given in § 5, to decide whether $x_t > 0$ on $[a, b]$ or $x_t < 0$ on $[a, b]$. A second test to decide between these two alternatives, based on likelihood ratios, is given in § 6. If the positive alternative is chosen then m_t^+ is used as an approximation to $\hat{x}_t$; and m_t^- is used if the negative alternative is chosen. Estimates for the error $\hat{x}_t - m_t^{\pm}$ are given in § 7. Finally, in § 8 extensions to multiple critical points of h and to dimension $n = l > 1$ are considered.

2. Assumptions: problem formulation. The following assumptions are made about the functions f, g, h in (1.1).

(A1) f, g, h are smooth with $g(x) > 0$. Moreover, f', g, g', g^{-1}, h', h'' are bounded on $-\infty < x < \infty$.

(A2) h has a single critical point, located at $x=0$, with $0=h(0)=h'(0)$ and $h(x)>0$ for $x\neq 0$. Moreover, $|h(x)|\to\infty$ as $|x|\to\infty$.

(A3) There exist y_1, y_2 with $0<y_1<y_2$ and $c>0$ such that

(2.1) $h(x^-)=h(x^+)=y$, $y_1\leqq y\leqq y_2$, and $x^-<0<x^+$ imply

$$|[g(x^-)h'(x^-)]^2-[g(x^+)h^-(x^+)]^2|>c.$$

Assumptions (A1) and (A2) imply that $\operatorname{sgn} h'(x)=\operatorname{sgn} x$, and that $\inf\{|h'(x)|: x\in K\}>0$ for any compact set $K\subset\Re^1-\{0\}$. Roughly speaking, the "detectability" condition (2.1) will be needed to distinguish whether $x_t>0$ or $x_t<0$ after observing y_s over a time interval during which no zero crossing of x_s has been detected. A sufficient condition for (2.1) is that $\psi'(x)\neq 0$ in some open interval containing $x=0$, where

$$\psi(x)=\frac{[g(x)h'(x)]^2}{h(x)}.$$

If $g(x)=$ constant, this sufficient condition holds provided $h''(x)>0$ and $h'''(0)\neq 0$. In [3] the case in which $h(x)$ is piecewise linear and $g(x)=$ constant is considered. In that case (2.1) is equivalent to the property that $h(-x)\neq h(x)$ for $x\neq 0$. The need for some condition like (2.1) is illustrated by the simple example $x_t=w_t$. If $h(-x)=h(x)$ for all x, the conditional distribution is always symmetric about $x=0$.

We formulate the nonlinear filter problem on the canonical sample space $\Omega=C(\Re_+)\times C(\Re_+)$, whose sample elements are denoted by (ω_1,ω_2) with $x_t(\omega)=\omega_1(t)$, $y_t(\omega)=\omega_2(t)$. Let $\mathcal{F}$ be the Borel field of Ω and P^ε a probability measure on $(\Omega,\mathcal{F})$ such that

$$x_t=x_0+\int_0^t f(x_s)\,ds+\int_0^t g(x_s)\,dw_s, \tag{2.2}$$

$$y_t=\frac{1}{\varepsilon}\int_0^t h(x_s)\,ds+v_t^\varepsilon \tag{2.3}$$

where w_t and v_t^ε are two mutually independent Brownian motions. The random variable x_0 is independent of $\{w_t, v_t^\varepsilon;\ t\geqq 0\}$ and for some $k>0$

$$E^\varepsilon[\exp(kx_0^2)]<\infty. \tag{2.4}$$

Note that y_t has the role of $\varepsilon^{-1}y_t^\varepsilon$ in (1.1). It is well known that P^ε exists and is unique (see, e.g., Stroock and Varadhan [17]). We define $\mathcal{Y}_t=\sigma\{y_s;\ 0\leqq s\leqq t\}$. The nonlinear filter problem is to find the conditional distribution μ_t of x_t given $\mathcal{Y}_t$. Our aim is to obtain an asymptotic result, as $\varepsilon\to 0$, concerning a finite-dimensional approximation to the nonlinear filter.

3. Approximations to $h(x_t)$. Let H be smooth for $-\infty<x<\infty$, with

$$0<c_1\leqq H'(x)\leqq c_2 \text{ and } H''(x) \text{ bounded}. \tag{3.1}$$

(In § 4 we shall take $H=h_+$, where h_+ is defined there.) Let m_t be the solution of

$$dm_t=f(m_t)\,dt+g(m_t)(dy_t-\varepsilon^{-1}H(m_t)\,dt) \tag{3.2}$$

with $m_0=E(x_0)$. Note that (3.2) has the same form as (1.2), except that h is replaced by H (recall that $H'(x)>0$ by (3.1)).

LEMMA 3.1. *Let* $0<a<b$. *Then for every* $\alpha>0$ *there exist positive* $\varepsilon_0=\varepsilon_0(\alpha)$, $K=K(\alpha)$, *such that*

$$P^\varepsilon\left(\sup_{[a,b]}|h(x_t)-H(m_t)|>\alpha\right)\leqq\exp\left(-\frac{K}{\varepsilon}\right),\qquad 0<\varepsilon<\varepsilon_0.$$

Proof. Let

$$L\phi=f(x)\phi'+\tfrac{1}{2}g^2(x)\phi'',$$

the generator of the Markov diffusion x_t. The Itô differential rule and (1.1), (3.2) imply

$$h(x_t)=h(x_0)+\int_0^t Lh(x_s)\,ds+\int_0^t (gh')(x_s)\,dw_s, \tag{3.3}$$

$$H(m_t)=H(m_0)+\int_0^t LH(m_s)\,ds+\int_0^t (gH')(m_s)(dy_s-\varepsilon^{-1}H(m_s))\,ds.$$

Let

$$z_t=h(x_t)-H(m_t),\qquad \rho_t=(gH')(m_t),$$
$$\zeta_t=Lh(x_t)-Lh(m_t),\qquad B_t=(w_t,v_t^\varepsilon),$$
$$\psi_t=((gh')(x_t),-(gH')(m_t)).$$

From (A1) and (3.1), $\rho_t\geqq\beta>0$. From (3.3) and (1.1)

$$z_t=z_0-\frac{1}{\varepsilon}\int_0^t\rho_s z_s\,ds+\int_0^t\zeta_s\,ds+\int_0^t\psi_s\cdot dB_s.$$

We write $z_t=z_t^{(1)}+z_t^{(2)}+z_t^{(3)}$, where

$$z_t^{(1)}=z_0\exp\left[-\frac{1}{\varepsilon}\int_0^t\rho_s\,ds\right],$$

$$z_t^{(2)}=\int_0^t\exp\left[-\frac{1}{\varepsilon}\int_s^t\rho_u\,du\right]\zeta_s\,ds,$$

$$z_t^{(3)}=\int_0^t\exp\left[-\frac{1}{\varepsilon}\int_s^t\rho_u\,du\right]\psi_s\cdot dB_s.$$

Then

$$|z_1^{(1)}|\leqq|z_0|\exp\left(-\frac{\beta t}{\varepsilon}\right),$$

$$P^\varepsilon\left(\sup_{[a,b]}|z_t^{(1)}|>\frac{\alpha}{3}\right)\leqq P^\varepsilon\left(|z_0|\geqq\frac{\alpha}{3}\exp\left(\frac{\beta a}{\varepsilon}\right)\right)\leqq\exp\left(-\frac{K_1}{\varepsilon}\right) \tag{3.4}$$

for $0<\varepsilon<\varepsilon_1(\alpha)$, using (2.4) and the Bienaymé-Chebyshev inequality. By the assumptions (A1) on f and g in § 2, there exists $k>0$ such that

$$E\exp(k\|\zeta\|_b)<\infty$$

where $\|\zeta\|_t=\sup_{[0,t]}|\zeta_s|$. Then

$$|z_t^{(2)}|\leqq\|\zeta\|_t\int_0^t\exp\left[-\frac{\beta(t-s)}{t}\right]ds.\qquad \|z^{(2)}\|_b\leqq\frac{\varepsilon}{\beta}\|\zeta\|_b,$$

$$P^\varepsilon\left(\|z^{(2)}\|_b>\frac{\alpha}{3}\right)\leqq P^\varepsilon\left(k\|\zeta\|_b>\frac{k\beta\alpha}{3\varepsilon}\right)\leqq\exp\left(-\frac{K_2}{\varepsilon}\right), \tag{3.5}$$

for $0<\varepsilon<\varepsilon_2(\alpha)$.

It remains to estimate $\|z^{(3)}\|_b$. Now $z_t^{(3)}$ satisfies the linear stochastic differential equation

$$dz_t^{(3)} = -\frac{1}{\varepsilon}\rho_t z_t^{(3)}\,dt + \psi_t \cdot dB_t, \qquad \rho_t \geqq \beta > 0$$

with $z_0^{(3)} = 0$. Moreover, by assumptions (A1) $|\psi_t|^2 \leqq k < \infty$. Define the linear operators L_t by

$$L_t F(z) = -\frac{1}{\varepsilon}\rho_t z F'(z) + \frac{1}{2}|\psi_t|^2 F''(z).$$

Let

$$F(z) = \exp\left(\frac{1}{2}\lambda z^2\right), \qquad \lambda = \frac{\beta}{k\varepsilon}.$$

Then

$$L_t F(z) = \left[\frac{1}{2}|\psi_t|^2(\lambda^2 z^2 + \lambda) - \frac{1}{\varepsilon}\lambda\rho_t z^2\right]F(z),$$

$$L_t F + \lambda F = \frac{1}{2}\lambda\left[2 + |\psi_t|^2 + \left(\lambda|\psi_t|^2 - \frac{2\rho_t}{\varepsilon}\right)z^2\right]F,$$

$$2 + |\psi_t|^2 + \left(\lambda|\psi_t|^2 - \frac{2\rho_t}{\varepsilon}\right)z^2 \leqq 2 + k - \frac{\beta}{\varepsilon}z^2,$$

since $\rho_t \geqq \beta$ and $\lambda k = \beta\varepsilon^{-1}$. Thus

$$L_t F + \lambda F \leqq 0 \quad \text{if } z^2 \geqq \frac{\varepsilon(2+k)}{\beta},$$

which implies

$$L_t F + \lambda F \leqq \frac{\lambda}{2}\max\left\{(2 + k + \lambda k z^2)F(z) : |z|^2 \leqq \frac{\varepsilon(2+k)}{\beta}\right\},$$

$$L_t F \leqq -\lambda F + N\lambda, \qquad N = \left(1 + \frac{k}{2}\right)\left(1 + \exp\left(\frac{2+k}{2k}\right)\right).$$

Following Kushner [10, pp. 79-80], for $0 \leqq t \leqq b$ let

$$W(z, t) = e^{\rho t}F(z) + \theta(e^{\rho b} - e^{\rho t})$$

where ρ and θ are to be chosen. Then

$$\frac{\partial W}{\partial t} + L_t W = e^{\rho t}[\rho F + L_t F - \theta\rho]$$

$$\leqq e^{\rho t}[(\rho - \lambda)F + N\lambda - \theta\rho].$$

We choose $\rho = \lambda^{1/2}$, $\theta = N\lambda^{1/2}$. Then for $\lambda \geqq 1$, $W(z_t^{(3)}, t)$ is a nonnegative supermartingale. Hence, for any $d > 0$

$$P^{\varepsilon}\left(\sup_{[0,b]} W(z_t^{(3)}, t) \geqq d\right) \leqq \frac{F(z_0) + \theta(e^{\rho b} - 1)}{\alpha}.$$

Since

$$\exp\left(\frac{1}{2}\lambda z^2\right) \leqq W(z, t) \quad \text{for } 0 \leqq t \leqq b \text{ and } F(z_0) = 1,$$

$$|z^{(3)}\|_b > \frac{\alpha}{3} \Rightarrow \sup_{[0,b]} W(z_t^{(3)}, t) \geqq \exp\left(\frac{\lambda\alpha^2}{18}\right),$$

$$P^\varepsilon\left(\|z^{(3)}\|_b > \frac{\alpha}{3}\right) \leqq \exp\left(-\frac{\lambda\alpha^2}{18}\right)[(1 + N\lambda^{1/2}b)\exp(\lambda^{1/2}b)].$$

Since $\lambda = \beta(k\varepsilon)^{-1}$ there exist K_3 and ε_3 such that

$$P^\varepsilon\left(\|z^{(3)}\|_b > \frac{\alpha}{3}\right) \leqq \exp\left(-\frac{K_3}{\varepsilon}\right), \qquad 0 < \varepsilon < \varepsilon_3. \tag{3.6}$$

From (3.4)-(3.6) we get Lemma 3.1.

4. Two possible approximate filters. We now define m_t^+ and m_t^- that satisfy equations like (1.2). A test will be given to determine whether x_t is positive and bounded away from zero on some time interval, with probability nearly one. When this test is positive, then m_t^+ will be used as an approximation to $\hat{x}_t$. Similarly, m_t^- will be used as an approximation to $\hat{x}_t$ when a corresponding test for negativity of x_t applies.

We first choose $\delta > 0$ such that

$$h(x) \geqq y_1 \quad \text{implies } |x| \geqq 2\delta \tag{4.1}$$

with $y_1 > 0$ as in (A3).

We also choose $y_3 > y_2$ and $r > 0$ (y_1 and y_2 are as in (A3)). Let

$$\bar{y}_1 = \sup_{|x| \leqq \delta} h(x), \qquad \bar{y}_3 = \inf_{|x| \geqq r} h(x).$$

By (A2), $\bar{y}_1 < y_1$ and $\bar{y}_3 > y_3$, provided we choose r large enough. The numbers y_3 and r may be regarded as large cutoffs for $h(x)$ and $|x|$, respectively. They will play a role only in § 7. In earlier sections, we can take $r = y_3 = \infty$.

We choose $h_+(x)$ and $h_-(x)$ such that

$$\begin{aligned} &h_+(x) = h(x) \quad \text{if } \delta \leqq x \leqq r, \\ &h_-(x) = h(x) \quad \text{if } -r \leqq x \leqq -\delta, \\ &c_1 \leqq h'_-(x) \leqq c_2, \qquad c_1 \leqq -h'_-(x) \leqq c_2 \end{aligned} \tag{4.2}$$

for all x, where $0 < c_1 < c_2$ and $h''_\pm(x)$ is bounded and continuous. Consider the following processes m_t^+, m_t^- for $t \geqq 0$:

$$dm_t^+ = f(m_t^+)\,dt + g(m_t^+)\left(dy_t - \frac{1}{\varepsilon}h_+(m_t^+)\,dt\right), \tag{4.3}$$

$$dm_t^- = f(m_t^-)\,dt - g(m_t^-)\left(dy_t - \frac{1}{\varepsilon}h(m_t^-)\,dt\right) \tag{4.4}$$

with initial data $m_0^\pm = E^\varepsilon(x_0)$. Note that $m_t^\pm$ depend on ε. Equations (4.3) and (4.4) are like (1.2) with h replaced by h_+ and h_-, respectively.

Lemma 3.1 provides a convenient test to determine whether x_t remains in one of the two intervals $[-r, -\delta]$ or $[\delta, r]$ during a time interval $[a, b]$. Define $B^{\pm} = B^{\pm}(a, b)$ and $C_\varepsilon^0 = C_\varepsilon^0(a, b)$ by

$$\begin{aligned} B^+ &= \{\delta \leqq x_t \leqq r \text{ for } a \leqq t \leqq b\}, \\ B^- &= \{-r \leqq x_t \leqq -\delta \text{ for } a \leqq t \leqq b\}, \\ C_\varepsilon^0 &= \{y_1 \leqq h_+(m_t^+) \leqq y_3 \text{ for } a \leqq t \leqq b\}. \end{aligned} \tag{4.5}$$

In Lemma 3.1 we now take $H = h_+$, $m_t = m_t^+$ and $\alpha < \min(y_1 - \bar{y}, \bar{y}_3 - y_3)$.

LEMMA 4.1. *Let $0 < a < b$. Then there exist positive ε_0 and K such that*

$$P^\varepsilon[(B^+ \cup B^-)^c \cap C_\varepsilon^0] \leqq \exp\left(-\frac{K}{\varepsilon}\right), \qquad 0 < \varepsilon < \varepsilon_0. \tag{4.6}$$

On B^+, $h_+(x_t) = h(x_t)$ for $a \leqq t \leqq b$. Since $h_+' \geqq c_1 > 0$ we also have from Lemma 3.1 the following. Given $\gamma > 0$ there exist ε_0 and K such that

$$P^\varepsilon\left(B^+ \cap \left\{\sup_{[a,b]} |x_t - m_t^+| > \delta\right\}\right) \leq \exp\left(-\frac{K}{\varepsilon}\right), \qquad 0 < \varepsilon < \varepsilon_0. \tag{4.7}$$

If we take α small enough in Lemma 3.1, then

$$P^\varepsilon(C_\varepsilon^0) \geqq P^\varepsilon(y_1 + \alpha \leqq h(x_t) \leqq y_3 - \alpha,\ a \leqq t \leqq b) - \exp\left(-\frac{K}{\varepsilon}\right).$$

Hence,

$$\liminf_{\varepsilon \to 0} P^\varepsilon(C_\varepsilon^0) > 0, \tag{4.8}$$

which implies an estimate like (4.6) for the conditional probability

$$P^\varepsilon[(B^+ \cup B^-)^c \,|\, C_\varepsilon^0].$$

5. Quadratic variation test. Next we let

$$C_\varepsilon = \{y_1 \leqq h_+(m_t^+) \leqq y_2 \text{ for } a \leqq t \leqq b\} \tag{5.1}$$

where y_1 and y_2 are as in (A3). Since $C_\varepsilon \subset C_\varepsilon^0$ we know from Lemma 4.1 that $P^\varepsilon[(B^+ \cup B^-)^c \cap C_\varepsilon]$ is exponentially small. Essentially the same proof as for (4.8) shows that $P^\varepsilon(C_\varepsilon) \geqq k > 0$ for small ε. Thus $P^\varepsilon[(B^+ \cup B^-)^c \,|\, C_\varepsilon]$ is also exponentially small. We shall introduce a test to decide between B^+ and B^-. This test is based on quadratic variations associated with the observation process y_t. It will be used in the course of justifying in § 6 another test of likelihood ratio type. Let $M = [\varepsilon^{-1}(b-a)]$ and for $j = 0, 1, \cdots, M-1$ define $t_j = a + j\varepsilon$

$$\begin{aligned} Y_j &= (y_{t_{j+1}} - y_{t_j}), \qquad S_j = \frac{1}{\varepsilon}\int_{t_j}^{t_{j+1}} h(x_s)\, ds, \\ V_j &= v_{t_{j+1}}^\varepsilon - v_{t_j}^\varepsilon. \end{aligned} \tag{5.2}$$

By (2.3), $Y_j = S_j + V_j$. If we consider only j even, then the random variables $V_{j+1} - V_j$ are independent. For the quadratic variation test, we need to estimate $\sum_j (Y_{j+1} - Y_j)^2$ for j even. For this purpose we first estimate $\sum_j (S_{j+1} - S_j)^2$ where the sum is taken over j even. We have

$$\begin{aligned} S_{j+1} - S_j &= \frac{1}{\varepsilon}\int_{t_j}^{t_{j+1}} [h(x_{s+\varepsilon}) - h(x_s)]\, ds \\ &= \frac{1}{\varepsilon}\int_{t_j}^{t_{j+1}} \int_s^{s+\varepsilon} p_\lambda\, dw_\lambda\, ds + \psi_j \end{aligned}$$

where $p_s = g(x_s)h'(x_s)$ and

$$\psi_j = \frac{1}{\varepsilon}\int_{t_j}^{t_{j+1}} \int_s^{s+\varepsilon} Lh(x_\lambda)\, d\lambda\, ds$$

with $Lh = f(x)h' + \frac{1}{2}g^2(x)h''$.

By assumption (A1)

$$|\psi_j| \leqq K\varepsilon \|x\|_b. \tag{5.3}$$

Let us introduce the "sawtooth" function ϕ such that, for j even,

$$\phi_s = \begin{cases} \dfrac{s - t_j}{\varepsilon} p_s, & s \in [t_j, t_{j+1}], \\ \dfrac{t_{j+2} - s}{\varepsilon} p_s, & s \in [t_{j+1}, t_{j+2}]. \end{cases}$$

By exchanging order of integration, we get

$$S_{j+1} - S_j = \int_{I_j} \phi_s\, dw_s + \psi_j \tag{5.4}$$

where $I_j = [t_j, t_{j+2}]$. Let

$$e = \sum_{j \text{ even}} \left[\left(\int_{I_j} \phi_s\, dw_s \right)^2 - \int_{I_j} \phi_s^2\, ds \right] \tag{5.5}$$

and define θ_s by

$$\theta_s = \int_{t_j}^{s} \phi_u\, dw_u, \qquad s \in I_j.$$

The Itô differential rule implies that

$$e = 2 \sum_{j \text{ even}} \int_{I_j} \theta_s \phi_s\, dw_s = 2 \int_a^b \theta_s \phi_s\, dw_s.$$

LEMMA 5.1. *Given $d > 0$ there exist positive ε_0 and K such that*

$$P^\varepsilon(|e| \geqq d) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad 0 < \varepsilon < \varepsilon_0.$$

Proof. We have

$$P^\varepsilon(|e| \geqq d) \leqq P^\varepsilon(\|\theta\| \geqq \varepsilon^{1/4}) + P^\varepsilon(|e| \geqq d, \|\theta\| < \varepsilon^{1/4})$$

where $\| \ \|$ is the sup-norm on $[a, b]$. Now

$$P^\varepsilon(\|\theta\| \geqq \varepsilon^{1/4}) = P^\varepsilon\left(\sup_{j \text{ even}} \sup_{I_j} |\theta_s| \geqq \varepsilon^{1/4} \right)$$

$$\leqq \sum_{j \text{ even}} P^\varepsilon\left(\sup_{I_j} |\theta_s| \geqq \varepsilon^{1/4} \right).$$

Since $|\phi_u^2| \leqq k$, a standard estimate (Theorem 18 of [15, p. 54]) gives

$$P^\varepsilon\left(\sup_{I_j} |\theta_s| \geqq c \right) \leqq 2 \exp\left(-\frac{c^2}{2k\varepsilon}\right).$$

We take $c = \varepsilon^{1/4}$ to get

$$P^\varepsilon(\|\theta\| \geqq \varepsilon^{1/4}) \leqq \frac{b-a}{\varepsilon} \exp\left(-\frac{1}{2k\varepsilon^{1/2}}\right).$$

By a slight modification of the argument in [15] just cited,

$$P^\varepsilon(|e| \geqq d, \|\theta\| < \varepsilon^{1/4}) \leqq 2 \exp\left(-\frac{d^2}{2k(b-a)\sqrt{\varepsilon}}\right).$$

We take $K < \min(1/2k, d^2/2k(b-a))$.

We now introduce Z_ε, on which the quadratic variation test will be based:

$$Z_\varepsilon = \frac{1}{b-a} \sum_{j \text{ even}} (Y_{j+1} - Y_j)^2. \tag{5.6}$$

LEMMA 5.2. *Given $d > 0$ there exist positive ε_0 and K such that*

$$P^\varepsilon\left(\left|Z_\varepsilon - \frac{1}{3(b-a)} \int_a^b (gh')^2(x_s)\, ds - 1\right| > d\right) \leqq \exp\left(-\frac{k}{\sqrt{\varepsilon}}\right)$$

for $0 < \varepsilon < \varepsilon_0$.

Remark. A similar result holds for *j odd*. Thus, for $\varepsilon < \varepsilon_0$,

$$P^\varepsilon\left(\left|\frac{1}{b-a}\left[\sum_{j=0}^{M-1} (Y_{j+1} - Y_j)^2 - \frac{2}{3}\int_a^b (gh')^2(x_s)\, ds\right] - 2\right| > 2d\right) \leqq 2 \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right).$$

This remark is useful when quadratic variations tests are implemented numerically [4].

Proof of Lemma 5.2. We write $Z_\varepsilon = \Gamma_1 + \Gamma_2 + \Gamma_3$, where

$$\Gamma_1 = \frac{1}{b-a} \sum_{j \text{ even}} (S_{j+1} - S_j)^2,$$

$$\Gamma_2 = \frac{2}{b-a} \sum_{j \text{ even}} (S_{j+1} - S_j)(V_{j+1} - V_j),$$

$$\Gamma_3 = \frac{1}{b-a} \sum_{j \text{ even}} (V_{j+1} - V_j)^2.$$

The lemma will be proved if we show that, for small ε

$$P^\varepsilon\left(\left|\Gamma_1 - \frac{1}{3(b-a)} \int_a^b (gh')^2(x_s)\, ds\right| > \frac{d}{3}\right) \leqq \exp\left(-\frac{K_1}{\sqrt{\varepsilon}}\right), \tag{5.7}$$

$$P^\varepsilon\left(|\Gamma_2| > \frac{d}{3}\right) \leqq \exp\left(-\frac{K_2}{\sqrt{\varepsilon}}\right), \tag{5.8}$$

$$P^\varepsilon\left(|\Gamma_3 - 1| > \frac{d}{3}\right) \leqq \exp\left(-\frac{K_3}{\sqrt{\varepsilon}}\right). \tag{5.9}$$

The random variables $V_{j+1} - V_j$, j even, $j = 0, 1, \cdots, M-1$, are independent and Gaussian, with the mean 0 and variance 2ε. Hence (5.9) follows from a standard large deviations theorem. We next recall (5.4) and write $\Gamma_1 = G_1 + G_2 + G_3$, where

$$G_1 = \frac{1}{b-a} \sum_{j \text{ even}} \left(\int_{I_j} \phi_s\, dw_s\right)^2,$$

$$G_2 = \frac{2}{b-a} \sum_{j \text{ even}} \psi_j \int_{I_j} \phi_s\, dw_s,$$

$$G_3 = \frac{1}{b-a} \sum_{j \text{ even}} \psi_j^2.$$

By standard estimates for stochastic differential equations, for fixed $\alpha > 0$ and small ε

$$P^\varepsilon\left(\sup_{|s-u|\leqq\varepsilon}|x_s - x_u| \geqq \alpha\right) \leqq \exp\left(-\frac{k}{\sqrt{\varepsilon}}\right).$$

Using Lemma 5.1, we then obtain

$$P^\varepsilon\left(\left|G_1 - \frac{1}{3(b-a)}\int_a^b (gh')^2(x_s)\,ds\right| > d_1\right) \leqq \exp\left(-\frac{K_1}{\sqrt{\varepsilon}}\right).$$

By (5.3), $|G_3| \leqq K\varepsilon\|x\|_b^2$, from which

$$P^\varepsilon(|G_3| > d_2) \leqq P\left(\|x\|_b > \left(\frac{d_2}{K\varepsilon}\right)^{1/2}\right) \leqq \exp\left(-\frac{k_2}{\sqrt{\varepsilon}}\right),$$

using the fact that $E\exp c\|x\|_b^2 < \infty$ for some $c < 0$. Finally, for any $\lambda > 0$,

$$|G_2| \leqq \lambda|G_1| + \frac{1}{\lambda}|G_3|,$$

$$P^\varepsilon(|G_2| > d_2) \leqq P^\varepsilon\left(\lambda|G_1| > \frac{d_2}{2}\right) + P^\varepsilon\left(\frac{1}{\lambda}|G_3| > \frac{d_2}{2}\right).$$

We then obtain (5.7) by choosing suitable $d_i < d/9$, $r = 1, 2, 3$ and λ sufficiently small.

To obtain (5.8) we first recall (5.3) and (5.4). Then

$$\begin{aligned} P^\varepsilon\left(\left|\sum_{j\text{ even}}\psi_j(V_{j+1} - V_j)\right| > \alpha\right) &\leqq P^\varepsilon\left(\frac{K\varepsilon M}{2}\|x\|_b \sup_{j\text{ even}}|V_{j+1} - V_j| > \alpha\right) \\ &\leqq P^\varepsilon\left(\frac{K(b-a)}{2\alpha}\|x\|_b > \varepsilon^{-1/4}\right) \\ &\quad + \left(\frac{M}{2}\right)P^\varepsilon(|V_{j+1} - V_j| > \varepsilon^{1/4}). \end{aligned}$$

At the last step we used mutual independence of the processes x_t, v_t^ε. By (2.4) and Theorem 5.7.2 of [7], $E^\varepsilon\exp[\sigma\|x\|_b^2] < \infty$ for some $\sigma > 0$. Thus,

$$P^\varepsilon\left(\left|\sum_{j\text{ even}}\psi_j(V_{j+1} - V_j)\right| > \alpha\right) \leqq \exp\left(-\frac{\gamma}{\sqrt{\varepsilon}}\right) + \frac{b-a}{2\varepsilon}\exp\left(-\frac{1}{8\sqrt{\varepsilon}}\right) \leqq \exp\left(-\frac{k_3}{\sqrt{\varepsilon}}\right),$$

for $k_3 < \min(\gamma, 1/8)$ and for small ε.

Next let us write

$$\sigma_s = \begin{cases} 1, & s \in [t_{j+1}, t_{j+2}), \\ -1, & s \in [t_j, t_{j+1}), \end{cases}$$

and consider the piecewise constant process ζ_s (independent of $\{V_j\}$) such that

$$\zeta_s = \int_{I_j}\phi_u\,dw_u \quad \text{for } s \in I_j.$$

Let $N_t = \int_a^t \zeta_s\sigma_s\,dv_s^\varepsilon$. Then N_t is a $\mathcal{F}_r$-martingale, and

$$N_b = \sum_{j\text{ even}}\left(\int_{I_j}\phi_u\,dw_u\right)(V_{j+1} - V_j),$$

$$\begin{aligned} \langle N\rangle_t &= \int_a^t \zeta_s^2\,ds = 2\varepsilon\sum_{k\text{ even}}\left(\int_{I_k}\phi_u\,dw_u\right)^2 \\ &= 2\varepsilon\int_a^b \phi_s^2\,ds + 2e\varepsilon. \end{aligned}$$

Since ϕ_s is bounded, by Lemma 5.1 we have for some Γ, $\Delta > 0$

$$P^\varepsilon(\langle N\rangle_t \geqq 2\Gamma\varepsilon) \leqq \exp\left(-\frac{\Delta}{\sqrt{\varepsilon}}\right).$$

We introduce the stopping time

$$\tau = b \wedge \inf\{t \in [a, b]: \langle N\rangle_t \geqq 2\Gamma\varepsilon\},$$

and let $N'_t = N_{t\wedge\tau}$. For any $\lambda > 0$

$$M_t = \exp\left(\lambda N'_t - \frac{\lambda^2}{2}\langle N'\rangle_t\right)$$

is a $\mathscr{F}_t$-martingale. Moreover, $\langle N\rangle_b \leqq 2\Gamma\varepsilon$ implies $N_t = N'_t$ for all $t \in [a, b]$. For $\alpha > 0$, $N_b > \alpha$, $\langle N\rangle_b \leqq 2\Gamma\varepsilon$ imply

$$M_b = \exp\left[\lambda N_b - \frac{\lambda^2}{2}\langle N\rangle_b\right] \geqq \exp(\lambda\alpha - \lambda^2\Gamma\varepsilon),$$

$$P^\varepsilon(N_b > \alpha, \langle N\rangle_b \leqq 2\Gamma\varepsilon) \leqq \exp(-\lambda\alpha + \lambda^2\Gamma\varepsilon).$$

We take $\lambda = \varepsilon^{-1}\theta$ with $\theta\Gamma < \alpha$, and conclude that for small ε

$$P^\varepsilon(N_b > \alpha) \leqq \exp\left(-\frac{k_4}{\sqrt{\varepsilon}}\right).$$

A similar argument, with λ replaced by $-\lambda$ gives, for small ε,

$$P^\varepsilon(N_b < -\alpha) \leqq \exp\left(-\frac{k_4}{\sqrt{\varepsilon}}\right).$$

We take $\alpha < d/6$ to obtain (5.8). This proves Lemma 5.2.

We shall now use assumption (A3) and Lemma 5.2 to describe a test for positivity or negativity of x_t on an interval $[a, b]$. As a function of y, $x^+(y)$ and $x^-(y)$ in (A3) are continuous. Under (A3), $(gh')^2(x^+) - (gh')^2(x^-)$ is continuous and not zero on the interval $[y_1, y_2]$. There are two possibilities: either

$$(gh')^2(x^-) < (gh')^2(x^+) - c \tag{5.10a}$$

for all $x^+ > 0$, $x^- < 0$ satisfying $h(x^+) = h(x^-) = y$, $y_1 \leqq y \leqq y_2$; or

$$(gh')^2(x^+) < (gh')^2(x^-) - c \tag{5.10b}$$

for all such x^+, x^-, y. We shall suppose that (5.10a) holds, the case of (5.10b) being entirely similar with a reversal of inequality signs in (5.11). Since x_s cannot itself be observed, we must approximate $(gh')^2(x_s)$ in Lemma 5.2 by quantities computable from an observation sample path. Let

$$\rho_s^+ = (gh'_+)(m_s^+), \qquad \rho_s^- = (gh'_-)(m_s^-),$$

and define $C_\varepsilon^\pm = C_\varepsilon^\pm(a, b)$ by

$$\begin{aligned} C_\varepsilon^+ &= C_\varepsilon \cap \left\{Z_\varepsilon > 1 + \frac{1}{6(b-a)}\int_a^b [(\rho_s^+)^2 + (\rho_s^-)^2]\, ds\right\}, \\ C_\varepsilon^- &= C_\varepsilon \cap \left\{Z_\varepsilon \leqq 1 + \frac{1}{6(b-a)}\int_a^b [(\rho_s^+)^2 + (\rho_s^-)^2]\, ds\right\}. \end{aligned} \tag{5.11}$$

Clearly $C_\varepsilon = C_\varepsilon^+ \cup C_\varepsilon^-$ and $C_\varepsilon^+ \cap C_\varepsilon^- = \varnothing$.

LEMMA 5.3. *There exist positive ε_0 and K such that, for $0<\varepsilon<\varepsilon_0$*

$$P^\varepsilon[(B^+)^c \cap C_\varepsilon^+] \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad P^\varepsilon[(B^-)^c \cap C_\varepsilon^-] \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right).$$

Proof. We prove the first inequality, the second being similar. First of all,

$$P^\varepsilon[(B^+)^c \cap C_\varepsilon^+] \leqq P^\varepsilon[(B^+ \cup B^-)^c \cap C_\varepsilon^0] + P^\varepsilon(B^- \cap C_\varepsilon^+).$$

The first term on the right side is estimated using Lemma 4.1. It remains to bound the last term. By (A1) and (5.10a) there exists $\alpha_0>0$ such that $\xi^+ \geqq 0$, $\xi^- <0$,

$$|h(x)-h(\xi^\pm)|<\alpha_0, \qquad y_1-\alpha_0 \leqq h(x) \leqq y_2+\alpha_0,$$

imply $\xi^+ \geqq \delta$, $\xi^- \leqq -\delta$ and

$$(gh')^2(\xi^-) < (gh')^2(\xi^+) - \frac{c}{2}.$$

Here δ is as in (4.1). Moreover, since gh' is bounded and Lipschitz

$$|(gh')^2(x)-(gh')^2(\xi^-)| \leqq K_1|x-\xi^-|.$$

Since $|h'(x)|$ is bounded away from zero for $|x| \geqq \delta$

$$|(gh')^2(x)-(gh')^2(\xi^-)| \leqq K_2|h(x)-h(\xi^-)|$$

for $x \leqq -\delta$, $\xi^- \leqq -\delta$.

We now take $x=x_s$, $\xi^\pm = m_s^\pm$. For $\alpha<\alpha_0$ consider the events

$$A_1 = \left\{(gh')^2(m_s^-) < (gh')^2(m_s^+) - \frac{c}{2}\right\},$$

$$A_2^+ = \{|(gh')^2(x_s)-(gh')^2(m_s^+)| < K_2\alpha\},$$

$$A_2^- = \{|(gh')^2(x_s)-(gh')^2(m_s^-)| < K_2\alpha\}.$$

We use Lemma 3.1 with $H=h_\pm$. For small ε

$$P^\varepsilon\left(\sup_{[a,b]}|h(x_t)-h_+(m_t^+)|>\alpha\right) \leqq \exp\left(-\frac{K}{\varepsilon}\right),$$

$$P^\varepsilon\left(\sup_{[a,b]}|h(x_t)-h_-(m_t^-)|>\alpha\right) \leqq \exp\left(-\frac{K}{\varepsilon}\right).$$

If $|h(x_t)-h_t(m_t^+)| \leqq \alpha$ and $|h(x_t)-h_-(m_t^-)| \leqq \alpha$, then

$$m_t^+ \geqq \delta, \qquad h_+(m_t^+) = h(m_t^+),$$

$$m_t^- \leqq -\delta, \qquad h_-(m_t^-) = h(m_t^-).$$

We have

$$B^- \cap C_\varepsilon \cap \left\{\sup_{[a,b]}|h(x_t)-h_\pm(m_t^\pm)| \leqq \alpha\right\} \subset A_1 \cap A_2^-,$$

$$P^\varepsilon[B^- \cap C_\varepsilon \cap (A_1 \cap A_2^-)^c] \leqq \exp\left(-\frac{K}{\varepsilon}\right)$$

for small ε.

We have on $B^- \cap C_\varepsilon \cap A_1 \cap A_2^-$,

$$\left| \int_a^b (gh')^2(x_s)\, ds - \int_a^b (\rho_s^-)^2\, ds \right| \leqq K_2(b-a)\alpha,$$

$$\int_a^b (\rho_s^-)^2\, ds < \int_a^b (\rho_s^+)^2\, ds - \frac{c}{2}(b-a),$$

$$1 + \frac{1}{3(b-a)} \int_a^b (gh')^2(x_s)\, ds < 1 + \frac{K_2\alpha}{3} + \frac{1}{3(b-a)} \int_a^b (\rho_s^-)^2\, ds$$

$$< 1 + \frac{K_2\alpha}{3} - \frac{c}{12} + \frac{1}{6(b-a)} \int_a^b [(\rho_s^+)^2 + (\rho_s^-)^2]\, ds.$$

We now take, in Lemma 5.2,

$$d = \frac{c}{12} - \frac{K_2\alpha}{3} > 0 \quad \text{if } \alpha < \frac{c}{4K_2},$$

to get for small ε

$$P^\varepsilon(B^- \cap C_\varepsilon^+ \cap A_1 \cap A_2^-) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right).$$

From these estimates we get Lemma 5.3.

The quadratic variation test now proceeds, roughly speaking, as follows. According to Lemma 4.1 the probability that x_t crosses zero during $[a, b]$ is exponentially small, if $C_\varepsilon^0(a, b)$. We will decide between $B^+(a, b)$ and $B^-(a, b)$ based on observing one of the two possible events $C_\varepsilon^+(a, b_1)$ or $C_\varepsilon^-(a, b_1)$, where $a < b_1 < b$. Thus, we let

$$Q_\varepsilon^+(a, b) = C_\varepsilon^0(a, b) \cap C_\varepsilon^+(a, b_1), \qquad Q_\varepsilon^-(a, b) = C_\varepsilon^0(a, b) \cap C_\varepsilon^-(a, b_1). \tag{5.12}$$

THEOREM 5.4. *There exist positive ε_0 and K such that, for $0 < \varepsilon < \varepsilon_0$:*

$$P^\varepsilon(B^+(a, b)^c \,|\, Q_\varepsilon^+(a, b)) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad P^\varepsilon(B^-(a, b)^c \,|\, Q_\varepsilon^-(a, b)) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right),$$

$$P^\varepsilon([B^+(a, b) \cup B^-(a, b)]^c \cap [Q_\varepsilon^+(a, b) \cup Q_\varepsilon^-(a, b)]) \leqq \exp\left(-\frac{K}{\varepsilon}\right).$$

Proof. We have

$$B^+(a, b)^c \cap Q_\varepsilon^+(a, b) \subset [B^+(a, b_1)^c \cap C_\varepsilon^+(a, b_1)] \cup [B^+(a, b_1) \setminus B^+(a, b) \cap C_\varepsilon^0(a, b)].$$

By Lemmas 4.1 and 5.3, for small ε and suitable K_1,

$$P^\varepsilon(B^+(a, b)^c \cap Q_\varepsilon^+(a, b)) \leqq \exp\left(-\frac{K_1}{\sqrt{\varepsilon}}\right). \tag{5.13}$$

By modifying arguments used in the proof of Lemma 5.3 we can easily show that

$$\liminf_{\varepsilon \to 0} P^\varepsilon(Q_\varepsilon^+(a, b)) > 0. \tag{5.14}$$

This implies the first inequality in Theorem 5.4, and the second inequality is similar. The third inequality is immediate from Lemma 4.1.

Remark. It can be seen that the constants K and ε_0 are uniform with respect to a, b_1, b, provided

$$0 < a^* \leqq a, \quad b \leqq b^* < \infty, \quad b - a \geqq t^* > 0,$$

where a^*, b^*, t^* are given. The result $C_\varepsilon^{\pm}(a, b_1)$ of testing on the interval $[a, b_1]$ is then used to decide whether to use the approximate filter m_t^+ or m_t^- up to time b. The accuracy of the approximate filter is discussed in § 7.

In this paper we consider the length $b_1 - a$ of the testing interval as fixed. In fact, an appropriate sequential decision test for positivity or negativity of x_t turns out to give a smaller mean decision time, for the same probabilities of incorrect decisions. This is discussed in [4], which considers a discrete time analogue of the model and in which some numerical results are reported.

6. Likelihood ratio test. Now we want to show how a decision between $B^+(a, b)$ and $B^-(a, b)$ can be made from the output m_t^+, m^- of the two approximate filters (4.4). For $a < d < b$ let us define the following test statistics, based on likelihood ratio considerations:

$$L_\varepsilon = \int_d^b (\hat{h}_s^+ - \hat{h}_s^-)\, dy_s - \frac{1}{2\varepsilon}\int_d^b [(\hat{h}_s^+)^2 - (\hat{h}_s^-)^2]\, ds \tag{6.1}$$

where $\hat{h}_s^+ = h_+(m_s^+)$, $\hat{h}_s^- = h_-(m_s^-)$. Using the representation

$$dy_s = \frac{1}{\varepsilon}\hat{h}_s\, ds + dv_s^\varepsilon \tag{6.2}$$

where $\hat{h}_s = E^\varepsilon[h(x_s)\,|\,\mathcal{Y}_s]$ and v_t^ε (the innovation) is a standard Brownian motion, L_ε can be rewritten in two ways. Let

$$z_s = \hat{h}_s^+ - \hat{h}_s^-.$$

Then

$$L_\varepsilon = \frac{1}{2\varepsilon}\int_d^b z_s^2\, ds + \int_d^b z_s\, dv_s^\varepsilon + \frac{1}{\varepsilon}\int_d^b z_s(\hat{h}_s - \hat{h}_s^+)\, ds, \tag{6.3$^+$}$$

$$L_\varepsilon = -\frac{1}{2\varepsilon}\int_d^b z_s^2\, ds + \int_d^b z_s\, dv_s^\varepsilon + \frac{1}{\varepsilon}\int_d^b z_s(\hat{h}_s - \hat{h}_s^-)\, ds. \tag{6.3$^-$}$$

We will show that the first term on the right side of (6.3$^\pm$) dominates the other two terms, except for events of small probability. In fact, this term is of order $O(1)$ (see Lemma 6.3). The sign of L_ε will then provide a test for positivity or negativity of x_t on $[a, b]$ (see Theorem 6.5).

Let us choose e with $a < e < d < b$. We need estimates for $\int_e^b |\hat{h}_s - \hat{h}_s^+|^2\, ds$. These are obtained in two steps, in Lemma 6.1 and 6.2. First, we replace $\hat{h}_s^+$ by a related quantity $\tilde{h}_s^+$ obtained after a change of probability measure. Second, we estimate $\hat{h}_s^+ - \tilde{h}_s^+$. Consider the following filtering problem:

$$dx_t = f(x_t)\, dt + g(x_t)\, dw_t,$$

$$dy_t = \frac{1}{\varepsilon}[h(x_t)1_{\{t \leqq e\}} + h_+(x_t)1_{\{t > e\}}]\, dt + dv_t^+.$$

The change of probability argument we now describe is due essentially to Ji. Define

$$\gamma_t = h_+(x_t) - h(x_t),$$

$$Z_t = \exp\left[\frac{1}{\varepsilon}\int_e^t \gamma_s\, dv_s^\varepsilon - \frac{1}{2\varepsilon^2}\int_e^t \gamma_s^2\, ds\right], \qquad e \leqq t \leqq b,$$

with $Z_t = 1$ for $0 \leqq t \leqq e$. Then v_t^+ is a P^+ standard Brownian motion, where

$$Z_b = \frac{dP^+}{dP^\varepsilon}.$$

Note that $P^+ = P^{+\varepsilon}$ depends on ε. Let us calculate $\hat{Z}_t = E(Z_t \mid \mathcal{Y}_t)$. We have

$$dZ_t = \frac{1}{\varepsilon}\gamma_t Z_t\, dv_t^\varepsilon, \qquad dy_t = \frac{1}{\varepsilon}h(x_t)\, dt + dv_t^\varepsilon.$$

The nonlinear filtering formula implies

$$d\hat{Z}_t = \frac{1}{\varepsilon}[\widehat{Z_t h_t} - \hat{Z}_t\hat{h}_t + \widehat{\gamma_t Z_t}]\, dv_t^\varepsilon$$

with $h_t = h(x_t)$ and with v_t^ε the innovation. Also

$$\widehat{Z_t\zeta_t} = \hat{Z}_t\tilde{\zeta}_t, \quad \text{where } \tilde{\zeta}_t = E^+(\zeta_t \mid \mathcal{Y}_t).$$

We take $\zeta_t = h_t$ and $\zeta_t = \gamma_t$ to get

$$d\hat{Z}_t = \frac{\hat{Z}_t}{\varepsilon}[\tilde{h}_t - \hat{h}_t + \tilde{\gamma}_t]\, dv_t^\varepsilon.$$

Moreover, $\tilde{\gamma}_t = \tilde{h}_t^+ - \tilde{h}_t$, where

$$\tilde{h}_t^+ = E^+(h_+(x_t) \mid \mathcal{Y}_t). \tag{6.4}$$

Hence,

$$d\hat{Z}_t = \frac{\hat{Z}_t}{\varepsilon}(\tilde{h}_t^+ - \hat{h}_t)\, dv_t^\varepsilon,$$

$$\log \hat{Z}_t = \int_e^t \left[\frac{1}{\varepsilon}(\tilde{h}_s^+ + \hat{h}_s)\, dv_s^\varepsilon - \frac{1}{2\varepsilon^2}(\tilde{h}_s^+ - \hat{h}_s)^2\, ds\right]. \tag{6.5}$$

By Jensen's inequality

$$-\log \hat{Z}_t \leqq -\widehat{\log Z_t}.$$

Since $\widehat{\log Z_t}$ and $\log \hat{Z}_t$ are continuous functions of t, and $C_\varepsilon^+(a, e)$ is $\mathcal{Y}_e$-measurable, we have that for any $\{\mathcal{Y}_t\}$-stopping time τ with $e \leqq \tau \leqq b$

$$E^\varepsilon[1_{C_\varepsilon^+(a,e)}\widehat{\log Z_\tau}] = E^\varepsilon[1_{C_\varepsilon^+(a,e)}\log Z_\tau],$$

$$E^\varepsilon\left[\frac{1}{2\varepsilon^2}\int_e^\tau (\tilde{h}_s^+ - \hat{h}_s)^2\, ds;\, C_\varepsilon^+(a, e)\right] = -E^\varepsilon[1_{C_\varepsilon^+(a,e)}\log \hat{Z}_\tau]$$

$$\leqq \frac{1}{2e^2}E^\varepsilon\left[\int_e^\tau (h_s^+ - h_s)^2\, ds;\, C_\varepsilon^+(a, e)\right].$$

We now take

$$\tau = b \wedge \inf\{t \geqq e\colon h_+(m_t^+) \notin [y_1, y_2]\}. \tag{6.6}$$

Then $h_s^+ = h_+(x_s) = h(x_s) = h_s$ on $B^+(e, \tau)$. By Cauchy-Schwartz

$$E^\varepsilon\left[\int_e^\tau (\tilde{h}_s^+ - \hat{h}_s)^2\, ds;\, C_\varepsilon^+(a, e)\right]$$

$$\leqq \left[E^\varepsilon\left(\int_a^b (h_s^+ - h_s)^2\, ds\right)^2\right]^{1/2}[P^\varepsilon(C_\varepsilon^+(a, e) \cap (B^+)^c(e, \tau))]^{1/2}$$

$$\leqq K[P^\varepsilon(C_\varepsilon^+(a, e) \cap (B^+)^c(e, \tau))]^{1/2}.$$

Also, on $B^+(a, e) \cap B^+(e, \tau)^c$,

$$\sup_{[e,b]} |h(x_t) - h_+(m_t^+)| > \alpha$$

for some $\alpha > 0$. We now apply Lemma 5.3 on the interval $[a, e]$ and Lemma 4.1 (or Lemma 3.1) on the interval $[e, b]$ to obtain Lemma 6.1.

LEMMA 6.1. *There exist positive $\varepsilon_0 > 0$ and K such that, for $0 < \varepsilon < \varepsilon_0$*

$$E^\varepsilon\left[\int_e^\tau (\tilde{h}_s^+ - \hat{h}_s)^2 \, ds;\ C_\varepsilon^+(a, e)\right] \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right).$$

We next estimate $\tilde{h}_s^+ - \hat{h}_s^+$ on $[e_1, b]$ where $e < e_1 < d < b$. We recall that m_t^+ as defined by (4.4) is an approximate filter of the form (1.2), when h is replaced by h_+ and v_t^ε by v_t^+. By (6.4) and Picard [12, formula (5.17)] there exist $\varepsilon_0 > 0$ and Γ_q (any $q \geqq 1$) such that for $\varepsilon < \varepsilon_0$

$$E^+[|\tilde{h}_s^+ - \hat{h}_s^+|^q] \leqq \Gamma_q \varepsilon^q, \qquad e_1 \leqq s \leqq b. \tag{6.7}$$

LEMMA 6.2. *For every $\alpha > 0$, $\mu > 0$, $q \geqq 1$,*

$$P^\varepsilon\left(\left\{\int_{e_1}^b |\tilde{h}_s^+ - \hat{h}_s^+|^q \, ds \geqq \alpha\varepsilon^\mu\right\} \cap B^+(a, b)\right) \leqq \frac{(b-a)\Gamma_q}{\alpha} \varepsilon^{q-\mu}.$$

Proof. We recall that $P^\varepsilon(D) = P^+(D)$ if $D \subset B^+(a, b)$. Lemma 6.2 then follows from (6.7).

Let us return to $z_s = \hat{h}_s^+ - \hat{h}_s^-$ in (6.3$^\pm$). From (4.2)-(4.4) and the Itô differential rule,

$$d\hat{h}_s^+ = (Lh_+)_s \, ds + \rho_s^+\left(\frac{\hat{h}_s - \hat{h}_s^+}{\varepsilon}\right) ds + \rho_s^+ \, d\nu_s^\varepsilon,$$

$$d\hat{h}_s^- = (Lh_-)_s \, ds - \rho_s^-\left(\frac{\hat{h}_s - \hat{h}_s^-}{\varepsilon}\right) ds - \rho_s^- \, d\nu_s^\varepsilon$$

where $(Lh_\pm)_s = Lh_\pm(m_s^\pm)$ and $\rho_s^\pm = (gh'_\pm)(m_s^\pm)$. Then

$$dz_s = [(Lh_+)_s - (Lh_-)_s] \, ds + \frac{\rho_s^-}{\varepsilon} z_s \, ds + \sigma_s \, d\nu_s^\varepsilon + \sigma_s\left(\frac{\hat{h}_s - \hat{h}_s^+}{\varepsilon}\right) ds, \tag{6.8}$$

$$\sigma_s = \rho_s^+ + \rho_s^-.$$

From assumptions (A1) and (4.2), $-\rho_s^- \geqq \beta > 0$. Moreover, by (5.10a), $x^+ > 0$, $x^- < 0$, $h(x^+) = h(x^-) \in [y_1, y_2]$ imply

$$(gh')(x^+) + (gh')(x^-) \geqq 2c_1 > 0,$$

for some $c_1 > 0$.

Using the notation of the proof of Lemma 5.3, we have

$$B^- \cap C_\varepsilon \cap \left\{\sup_{[a,b]} |h(x_t) - h_-(m_t^-)| \leqq \alpha\right\} \subset A_2^-,$$

$$B^+ \cap C_\varepsilon \cap \left\{\sup_{[a,b]} |h(x_t) - h_+(m_t^+)| \leqq \alpha\right\} \subset A_2^+.$$

By using Lemma 3.1 in the same way as in the proof of Lemma 5.3, together with Lemma 4.1, we have

$$P^\varepsilon\left(\left\{\min_{[a,b]}\sigma_s < c_1\right\}\cap C_\varepsilon(a,b)\right)\leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right) \tag{6.9}$$

for small ε.

We now define

$$A_\varepsilon^+ = C_\varepsilon^+(a,e)\cap C_\varepsilon(a,b),\quad A_\varepsilon^- = C_\varepsilon^-(a,e)\cap C_\varepsilon(a,b). \tag{6.10}$$

LEMMA 6.3. *There exist positive* γ, ε_m, K_m, $m=1,2,\cdots$, *such that for* $0<\varepsilon<\varepsilon_m$

$$P^\varepsilon\left(\left\{\frac{1}{\varepsilon}\int_d^b z_s^2\,ds<\gamma\right\}\cap A_\varepsilon^+\right)\leqq K_m\varepsilon^m.$$

Proof. We write $z_s = z_s^{(1)}+z_s^{(2)}$, where $z_{e_1}^{(1)}=0$ and (see (6.8))

$$dz_s^{(1)} = \frac{\rho_s^- z_s^{(1)}}{\varepsilon}\,ds+\sigma_s\,d\nu_s^\varepsilon,$$

$$dz_s^{(2)} = \frac{\rho_s^- z_s^{(2)}}{\varepsilon}\,ds+[(Lh_+)_s-(Lh_-)_s]\,ds+\sigma_s\left(\frac{\hat h_s-\hat h_s^+}{\varepsilon}\right)ds.$$

We first show that the contribution of $z_s^{(2)}$ is small. We have

$$z_t^{(2)} = z_{e_1}\exp\int_{e_1}^t\frac{\rho_s\,ds}{\varepsilon}+\int_{e_1}^t\exp\left[\int_s^t\frac{\rho_u^-\,du}{\varepsilon}\right][(Lh_+)_s-(Lh_-)_s]\,ds+\xi_t+\eta_t,$$

$$\xi_t = \int_{e_1}^t\exp\left[\int_s^t\frac{\rho_u^-\,du}{\varepsilon}\right]\sigma_s\frac{(\hat h_s-\tilde h_s^+)}{\varepsilon}\,ds,$$

$$\eta_t = \int_{e_1}^t\exp\left[\int_s^t\frac{\rho_u^-\,du}{\varepsilon}\right]\sigma_s\frac{(\tilde h_s^+-\hat h_s^+)}{\varepsilon}\,ds.$$

Since $\rho_s^-\leqq-\beta<0$, the first term is exponentially small on $[d,b]$. Since $(Lh_\pm)_s$ is bounded, the second term is of order ε. An argument similar to one in the proof of Lemma 3.1 shows that it suffices to estimate ξ_t and η_t. Using $\rho_u^-\leqq-\beta<0$, σ_s bounded and Cauchy-Schwartz

$$\xi_t^2\leqq N\int_s^t\frac{(\hat h_s-\tilde h_s^+)^2\,ds}{\varepsilon}.$$

Now $A_\varepsilon^+\subset C_\varepsilon^+(a,e)$ and $\tau=b$ on A_ε^+. By Lemma 6.1, given $\theta>0$

$$P^\varepsilon\left(\left\{\frac{1}{\varepsilon}\sup_{[e_1,b]}\xi_t^2\geqq 0\right\}\cap A_\varepsilon^+\right)\leqq P^\varepsilon\left(\left\{\int_{e_1}^\tau(\hat h_s-\tilde h_s^+)^2\,ds\geqq\frac{\theta\varepsilon^2}{N}\right\}\cap C_\varepsilon^+(a,e)\right)$$

$$\leqq\frac{N}{\theta\varepsilon^2}\exp\left(-\frac{K}{\sqrt{\varepsilon}}\right)<\varepsilon^m$$

for small ε (depending on m).

To estimate η_t^2 we use Hölder's inequality with $1<p<2$, $p^{-1}+q^{-1}=1$:

$$\eta_t^2\leqq\frac{k_1}{\varepsilon^2}\left(\int_{e_1}^t\exp\left[\frac{p}{\varepsilon}\int_s^t\rho_u^-\,du\right]ds\right)^{2/p}\left(\int_{e_1}^t|\tilde h_s^+-\hat h_s^+|^q\,ds\right)^{2/q}.$$

Since $\rho_s^-\leqq-\beta<0$,

$$\frac{1}{\varepsilon}\eta_t^2 \leqq k_2 \varepsilon^{2/p-3}\left(\int_{e_1}^{t} |\tilde{h}_s^+ - \hat{h}_s^+|^q \, ds\right)^{2/q},$$

$$P^\varepsilon\left(\left\{\frac{1}{\varepsilon}\sup_{[e,b]} \eta_t^2 \geqq \theta\right\} \cap A_\varepsilon^+\right) \leqq P^\varepsilon(B^+(a,b)^c \cap A_\varepsilon^+)$$

$$+ P^\varepsilon\left(\left\{\int_{e_1}^{b} |\tilde{h}_s^+ - \hat{h}_s^+|^q \, ds \geqq \alpha\varepsilon^\mu\right\} \cap B^+(a,b)\right),$$

with $\alpha = k_2^{-q/2}$ and

$$\mu = \left(3 - \frac{2}{p}\right)\left(\frac{q}{2}\right) = \frac{1}{2} + 1.$$

The same proof as for Lemma 5.3 shows that

$$P^\varepsilon(B^+(a,b)^c \cap A_\varepsilon^+) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad 0 < \varepsilon < \varepsilon_0,$$

with a similar inequality for B^- and A_ε^-.

By Lemma 6.2 we then have

$$P^\varepsilon\left(\left\{\frac{1}{\varepsilon}\sup_{[e,b]} \eta_t^2 \geqq \theta\right\} \cap A_\varepsilon^+\right) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right) + \frac{(b-a)\Gamma_q}{\alpha}\varepsilon^{q/2-1} < \varepsilon^m$$

for small ε, if we choose $q > 2m+2$.

It remains to consider $z_s^{(1)}$. Let

$$\begin{aligned} &u = \varepsilon^{-1}(s - e_1), \qquad \zeta_u = \varepsilon^{-1/2} z_s^{(1)}, \\ &\phi_u = -\rho_s^-, \quad \Sigma_u = \sigma_s, \quad V_u = \varepsilon^{-1/2}\nu_s^\varepsilon. \end{aligned} \tag{6.11}$$

Then V_u is a standard Brownian motion and

$$d\zeta_u = -\phi_u \zeta_u \, du + \Sigma_u \, dV_u, \qquad u \geqq 0, \quad \zeta_0 = 0. \tag{6.12}$$

Moreover,

$$\frac{1}{\varepsilon}\int_d^b [z_s^{(1)}]^2 \, ds = \frac{b-e_1}{U}\int_{U_1}^{U} \zeta_u^2 \, du \quad \text{where } U_1 = \frac{d-e_1}{\varepsilon}, \quad U = \frac{b-e_1}{\varepsilon}.$$

Since $\phi_u \geqq \beta > 0$, to complete the proof of Lemma 6.3 it will suffice to show that, for some $\bar{c} > 0$, there exist $\bar{\varepsilon}_m$, $\bar{K}_m$ such that

$$P^\varepsilon\left(\left\{\frac{1}{U}\int_{U_1}^{U} \phi_u \zeta_u^2 \, du < \bar{c}\right\} \cap A_\varepsilon^+\right) \leqq \bar{K}_m \varepsilon^m, \qquad 0 < \varepsilon < \bar{\varepsilon}_m. \tag{6.13}$$

We write $\zeta_u^2 = \psi_u + \bar{\psi}_u$, where from (6.12) and the Itô differential rule

$$\begin{aligned} d\psi_u &= \left(-2\phi_u\psi_u + \Sigma_u^2\right) du, \qquad \psi_0 = 0, \\ d\bar{\psi}_u &= -2\phi_u\bar{\psi}_u + 2\,\Sigma_u\,\zeta_u \, dV_u, \qquad \bar{\psi}_0 = 0. \end{aligned}$$

Let $\bar{A}_\varepsilon^+ = A_\varepsilon^+ \cap \{\sum_u \geqq c_1\}$, with c_1 as in (6.9).

For small ε

$$P^\varepsilon(A_\varepsilon^+ \setminus \bar{A}_\varepsilon^+) \leqq \exp\left(-\frac{K}{\sqrt{\varepsilon}}\right). \tag{6.14}$$

Since $0 \leqq \beta \leqq \phi_u \leqq M < \infty$ and $\sum_u$ is bounded, elementary estimates imply that on $\bar{A}_\varepsilon^+$

$$\psi_u \geqq \frac{c_1^2}{2M}(1 - e^{-2Mu}) \geqq \frac{c_1^2}{3M}$$

for $u \geqq U_1$ and small enough ε. Thus, on $\bar{A}_\varepsilon^+$

$$\frac{1}{U}\int_{U_1}^{U} \phi_u \psi_u \, du \geqq \frac{c_1^2 \beta}{3M}\left(1 - \frac{U}{U_1}\right) \geqq c_2 > 0 \quad \text{since } \frac{U_1}{U} = \frac{d - e_1}{b - e_1}. \tag{6.15}$$

In (6.13) we take $\bar{c} = \frac{1}{2}c_2$.

On the other hand,

$$\bar{\psi}_u = -2\int_0^u \phi_v \bar{\psi}_v \, dv + M_u \quad \text{where } M_u = 2\int_0^u \Sigma_v \, \zeta_v \, dV_v.$$

Since ζ_u satisfies (6.12) with $\sum_u$ bounded and $\phi_u \geqq \beta > 0$, standard arguments imply that $E|\zeta_u|^l$ is bounded on $0 \leqq u < \infty$, for each $l \geqq 1$. Since $\sum_u$ is bounded and M_u is a martingale, this implies

$$E\|M\|_U^{2m} \leqq \Lambda_m U^m, \qquad m = 1, 2, \cdots,$$

for suitable Λ_m. Since $\tilde{\psi}_u = \zeta_u^2 - \psi_u$, and ψ_u is bounded, by taking $l = 4m$ we see that $E|\tilde{\psi}_u|^{2m}$ is bounded for each m. Then, for small ε,

$$P^\varepsilon\left(\frac{1}{U}\left|\int_{U_1}^{U} \phi_v \tilde{\psi}_v \, dv\right| > \frac{\bar{c}}{2}\right) \leqq P^\varepsilon\left(\frac{1}{U}(|\tilde{\psi}_U| + |\tilde{\psi}_{U_1}|) > \frac{\bar{c}}{8}\right) + P^\varepsilon\left(\frac{1}{U}|M_{U_1}| > \frac{\bar{c}}{8}\right) \leqq \bar{\Lambda}_m \varepsilon^m,$$

using Chebyshev's inequality. From this, together with (6.14) and (6.15), we obtain (6.13) and hence Lemma 6.3.

LEMMA 6.4. *There exist positive* ε_m, K_m, $m = 1, 2, \cdots$, *such that for* $0 < \varepsilon < \varepsilon_m$:

$$P^\varepsilon(\{L_\varepsilon < 0\} \cap A_\varepsilon^+) \leqq K_m \varepsilon^m, \qquad P^\varepsilon(\{L_\varepsilon \geqq 0\} \cap A_\varepsilon^-) \leqq K_m \varepsilon^m.$$

Proof. Let us prove the first estimate only. Using (6.3^+) we rewrite L_ε as

$$L_\varepsilon = \left[\frac{1}{4\varepsilon}\int_d^b z_s^2 \, ds + \int_d^b z_s \, d\nu_s\right] + \left[\frac{1}{4\varepsilon}\int_d^b z_s^2 \, ds + \frac{1}{\varepsilon}\int_d^b z_s(\hat{h}_s - \hat{h}_s^+) \, ds\right].$$

Let us first show that the sum of the last two terms is nonnegative on A_ε^+ with probability nearly one. Indeed, it is bounded below by

$$X = \frac{1}{8\varepsilon}\int_d^b z_s^2 \, ds - \frac{2}{\varepsilon}\int_d^b (\hat{h}_s - \hat{h}_s^+)^2 \, ds.$$

The same arguments used in the proof of Lemma 6.3 to estimate ξ_t and η_t show that

$$P^\varepsilon\left(\left\{\frac{16}{\varepsilon}\int_d^b (\hat{h}_s - \tilde{h}_s^+)^2 \, ds > \frac{\gamma}{2}\right\} \cap A_\varepsilon^+\right) \leqq \bar{K}_m \varepsilon^m$$

and hence from Lemma 6.3,

$$P^\varepsilon(\{X < 0\} \cap A_\varepsilon^+) \leqq K^{(2)} \varepsilon^m$$

for small ε. Let

$$M_t = \varepsilon^{-1/2}\int_d^t z_s \, d\nu_s^\varepsilon.$$

The sum of the first two terms is negative provided that

$$M_b + \frac{1}{4\sqrt{\varepsilon}}\langle M\rangle_b < 0.$$

Let $\alpha < \gamma$ with γ as in Lemma 6.3, and let

$$D = \left\{ M_b < -\frac{1}{4\sqrt{\varepsilon}}\langle M\rangle_b \right\} \cap \{M_b \geqq \alpha\}.$$

For $\lambda < 0$ we have on D

$$\lambda M_b - \frac{\lambda^2}{2}\langle M\rangle_b \geqq \left(-\frac{\lambda}{4\sqrt{\varepsilon}} - \frac{\lambda^2}{2}\right)\langle M\rangle_b \geqq \alpha\left(-\frac{\lambda}{4\sqrt{\varepsilon}} - \frac{\lambda^2}{2}\right).$$

Since $E^\varepsilon[\exp(\lambda M_b - \frac{1}{2}\lambda^2\langle M\rangle_b)] = 1$, we have by choosing λ suitable that for any $k > 0$ and small ε,

$$P^\varepsilon(D) \leqq \exp\left(-\frac{k}{\sqrt{\varepsilon}}\right).$$

This completes the proof of Lemma 6.4.

From (5.11) and (6.10), $C_\varepsilon = A_\varepsilon^+ \cup A_\varepsilon^-$ and $A_\varepsilon^+ \cap A_\varepsilon^- = \varnothing$. We then conclude from Lemmas 4.1 and 6.4, together with (6.11), the following corollary.

COROLLARY 6.5. *There exist positive* ε_m, K_m, $m = 1, 2, \cdots$, *such that for* $0 < \varepsilon < \varepsilon_m$,

$$P^\varepsilon(\{L_\varepsilon < 0\} \cap C_\varepsilon(a, b) \cap B^+(a, b)) < K_m\varepsilon^m,$$
$$P^\varepsilon(\{L_\varepsilon \geqq 0\} \cap C_\varepsilon(a, b) \cap B^-(a, b)) < K_m\varepsilon^m.$$

In much the same way as for the test $Q_\pm^\varepsilon$ at the end of § 5, we now introduce a test $R_\pm^\varepsilon$ based on L_ε. In (6.1) let us now write $L_\varepsilon = L_\varepsilon(d, b)$. We consider $a < d < b_1 < b$, and let

$$\begin{aligned} R_\varepsilon^+(a, b) &= C_\varepsilon^0(a, b) \cap \{L_\varepsilon(d, b_1) \geqq 0\} \cap C_\varepsilon(a, b_1), \\ R_\varepsilon^-(a, b) &= C_\varepsilon^0(a, b) \cap \{L_\varepsilon(d, b_1) < 0\} \cap C_\varepsilon(a, b_1). \end{aligned} \tag{6.16}$$

From Lemma 4.1 and Corollary 6.5, we obtain, by the same proof as for Theorem 5.4, the following theorem.

THEOREM 6.6. *There exist positive* K, ε_m, K_m *for* $m = 1, 2, \cdots$, *such that for* $0 < \varepsilon < \varepsilon_m$,

$$P^\varepsilon(B^+(a, b)^c \,|\, R_\varepsilon^+(a, b)) \leqq K_m\varepsilon^m, \qquad P^\varepsilon(B^-(a, b)^c \,|\, R_\varepsilon^-(a, b)) \leqq K_m\varepsilon^m,$$
$$P^\varepsilon([B^+(a, b) \cup B^-(a, b)]^c \cap [R_\varepsilon^+(a, b) \cup R_\varepsilon^-(a, b)]) \leqq \exp\left(-\frac{K}{\varepsilon}\right).$$

7. Approximations to the conditional mean. In §§ 5 and 6 we described two tests to decide between the events $B^+(a, b)$ and $B^-(a, b)$, with small probability of error when ε is small. It remains to estimate how well the conditional mean $\hat{x}_t$ is approximated by m_t^+ in the positive case or by m_t^- in the negative case. Let us consider the events $Q_\varepsilon^\pm(a, b)$ defined by (5.12), using the quadratic variations test. The discussion with $R_\varepsilon^\pm(a, b)$ defined by (6.16) is similar, except that estimates exponential in ε are replaced by estimates polynomial in ε.

For brevity we write $B^\pm = B^\pm(a, b)$, $Q_\varepsilon^\pm = Q_\varepsilon^\pm(a, b)$, etc.

LEMMA 7.1. *For each* $q > 0$ *there exist positive* ε_q, K_q *such that, for* $0 < \varepsilon < \varepsilon_q$

$$E^\varepsilon[|\hat{x}_b - m_b^+|^q;\ Q_\varepsilon^+ \cap (B^+)^c] \leqq \exp\left(-\frac{K_q}{\sqrt{\varepsilon}}\right).$$

Proof. By Cauchy-Schwartz,

$$\text{leftside} \leqq (E^\varepsilon|\hat{x}_b - m_b^+|^{2q})^{1/2}(P^\varepsilon(Q_\varepsilon^+ \cap (B^+)^c))^{1/2}.$$

The last term is estimated by (5.13). Since $E^\varepsilon|\hat{x}_b|^{2q}$ is bounded, it remains only to bound $E^\varepsilon|m_b^+|^{2q}$. Let $z_t = h(x_t) - h_+(m_t^+)$. Then as in the proof of Lemma 3.1,

$$dz_t = -\frac{1}{\varepsilon}\rho_t z_t\,dt + \zeta_t\,dt + \psi_t \cdot dB_t$$

with $\rho_t \geqq \beta > 0$, ζ_t and ψ_t bounded, and $B_t = (w_t, v_t^\varepsilon)$ a two-dimensional Brownian motion. This implies a bound on $E^\varepsilon|z_b|^{2q}$; in fact it is of order ε^q. Since $E^\varepsilon|h(x_b)|^{2q}$ is bounded and $|m_b^+| \leqq k_1|h_+(m_b^+)| + k_2$, we obtain the required bound for $E^\varepsilon|m_b^+|^{2q}$. This proves Lemma 7.1.

We next make a change of probability measure, from P^ε to P^+ with

$$\frac{dP^+}{dP^\varepsilon} = Z_b, \qquad Z_t = \exp\left[\frac{1}{\varepsilon}\int_a^t \gamma_s\,dv_s^\varepsilon - \frac{1}{2\varepsilon^2}\int_a^t \gamma_s^2\,ds\right],$$

$$\gamma_s = h_+(x_s) - h(x_s).$$

This is the same as in § 6, except that the lower limit e is replaced by a.

LEMMA 7.2. *There exist positive ε_0, K such that*

$$P^+((B^+)^c \cap C_\varepsilon^0) \leqq \exp\left(-\frac{K}{\varepsilon}\right), \qquad 0 < \varepsilon < \varepsilon_0.$$

This is proved in the same way as Lemma 4.1 with h replaced by h_+, recalling (4.1) and (4.2).

We use the notation $\phi_t = \phi(x_t)$, and

$$\hat{\phi}_t = E^\varepsilon[\phi_t \mid \mathcal{Y}_t], \qquad \tilde{\phi}_t = E^+[\phi_t \mid \mathcal{Y}_t]. \tag{7.1}$$

As already used in the proof of Lemma 6.1,

$$\widehat{Z_t\phi_t} = \hat{Z}_t\tilde{\phi}_t.$$

LEMMA 7.3. *Let ϕ be bounded and continuous and $\alpha > 0$. Then, for every $q > 0$, there exist positive ε_q, K_q such that*

$$P^\varepsilon(\{|\widehat{\phi_b Z_b} - \hat{\phi}_b| > \alpha\varepsilon^q\} \cap Q_\varepsilon^+) \leqq \exp\left(-\frac{K_q}{\sqrt{\varepsilon}}\right)$$

for $0 < \varepsilon < \varepsilon_0$.

Proof. We recall that $P^\varepsilon(E) = P^+(E)$ for $E \subset B^+$ and that $Z_b = 1$ on B^+. If D is $\mathcal{Y}_b$-measurable, then

$$\int_D (\widehat{\phi_b Z_b} - \hat{\phi}_b)\,dP^\varepsilon = \int_D \phi_b Z_b\,dP^\varepsilon - \int_D \phi_b\,dP^\varepsilon$$

$$= \int_{D\cap(B^+)^c} \phi_b\,dP^+ - \int_{D\cap(B^+)^c} \phi_b\,dP^\varepsilon.$$

Let

$$D = \{\widehat{\phi_b Z_b} - \hat{\phi}_b > \alpha\varepsilon^q\} \cap Q_\varepsilon^+.$$

Using (5.13), Lemma 7.2 and $Q_\varepsilon^+ \subset C_\varepsilon^0$, we have for small ε and suitable K_q

$$P^\varepsilon(\{\widehat{\phi_b Z}_b - \hat{\phi}_b > \alpha\varepsilon^q\} \cap Q_\varepsilon^+) \leqq \frac{\|\phi\|}{\alpha\varepsilon^q}[P^+((B^+)^c \cap Q_\varepsilon^+) + P^\varepsilon((B^+)^c \cap Q_\varepsilon^+)]$$

$$\leqq \frac{1}{2}\exp\left(-\frac{K_q}{\sqrt{\varepsilon}}\right).$$

Similarly, for small ε,

$$P^\varepsilon(\{\widehat{\phi_b Z}_b - \hat{\phi}_b < -\alpha\varepsilon^q\} \cap Q_\varepsilon^+) \leqq \frac{1}{2}\exp\left(-\frac{K_q}{\sqrt{\varepsilon}}\right).$$

This proves Lemma 7.3.

Lemma 7.3 holds in particular for $\phi_t = 1$. Thus, for small ε,

$$P^\varepsilon(\{|\hat{Z}_b - 1| > \alpha\varepsilon^q\} \cap Q_\varepsilon^+) \leqq \exp\left(-\frac{K_q}{\sqrt{\varepsilon}}\right).$$

We then conclude from Lemma 7.3 that given $\theta > 0$ we have for suitable $\bar{\varepsilon}_q$, $\bar{K}_q$

$$P^\varepsilon(\{|\tilde{\phi}_b - \hat{\phi}_b| > \theta\varepsilon^q\} \cap Q_\varepsilon^+) \leqq \exp\left(-\frac{\bar{K}_q}{\sqrt{\varepsilon}}\right) \tag{7.2}$$

for $0 < \varepsilon < \bar{\varepsilon}_q$.

With r as in § 4, we choose ϕ bounded and continuous such that $\phi(x) = x$ if $|x| \leqq r$ and $|\phi(x)| \leqq |x|$ for all x.

LEMMA 7.4. *There exist positive ε_0 and K such that, for $0 < \varepsilon < \varepsilon_0$, and any $D \subset Q_\varepsilon^+$ which is $\mathcal{Y}_b$-measurable*

$$\left|\int_D (\hat{\phi}_b - \hat{x}_b)\, dP^\varepsilon\right| \leqq \exp\left(-\frac{K}{\varepsilon}\right), \qquad \left|\int_D (\tilde{\phi}_b - \tilde{x}_b)\, dP^+\right| \leqq \exp\left(-\frac{K}{\varepsilon}\right).$$

Proof. Let $B_r = \{|x_t| \leqq r \text{ for } a \leqq t \leqq b\}$. Then

$$\left|\int_D (\hat{\phi}_b - \hat{x}_b)\, dP^\varepsilon\right| = \left|\int_D (\phi_b - X_b)\, dP^\varepsilon\right|$$

$$\leqq (E\phi_b^2 + Ex_b^2)^{1/2}[P^\varepsilon(Q_\varepsilon^+ \cap B_r^c)]^{1/2}.$$

The first inequality then follows from Lemma 4.1, since $Q_\varepsilon^+ \subset C_\varepsilon^0$ and $B_r^c \subset (B^+ \cup B^-)^c$. The second inequality is proved in the same way, using Lemma 7.2 and $B_r^c \subset (B^+)^c$. Next, we recall from Picard [12] that for any $q > 0$ there exists N_q, such that for small ε

$$E^+[|\tilde{x}_b - m_b^+|^q] \leqq N_q\varepsilon^q. \tag{7.3}$$

THEOREM 7.5. *Let $0 < p < 1$. Then, for $l = 1, 2, \cdots$ there exist ε_l, K_l such that*

$$P^\varepsilon(|\hat{x}_b - m_b^+| > \varepsilon^p \,|\, Q_\varepsilon^+) \leqq K_l\varepsilon^l,$$

$$P^\varepsilon(|\hat{x}_b - m_b^-| > \varepsilon^p \,|\, Q_\varepsilon^-) \leqq K_l\varepsilon^l, \qquad 0 < \varepsilon < \varepsilon_l.$$

Proof. It suffices to prove the first inequality. We have

$$P^\varepsilon(|\hat{x}_b - m_b^+| > \varepsilon^p;\, Q_\varepsilon^+) \leqq P^\varepsilon((B^+)^c \cap Q_\varepsilon^+) + P^\varepsilon(|\hat{x}_b - m_b^+| > \varepsilon^p;\, B^+ \cap Q_\varepsilon^+).$$

The first term on the right side is estimated from (5.13). To estimate the second term, we write

$$\bar{x}_b - m_b^+ = \hat{x}_b - \tilde{x}_b + \tilde{x}_b - m_b^+.$$

Let us choose ϕ as in Lemma 7.4. Then

$$P^\varepsilon(\hat{x}_b - \tilde{x}_b > \tfrac{1}{2}\varepsilon^p; Q_\varepsilon^+) \leqq P^\varepsilon(|\hat{\phi}_b - \tilde{\phi}_b| > \tfrac{1}{6}\varepsilon^p; Q_\varepsilon^+) + P^\varepsilon(\hat{\phi}_b - \hat{x}_b > \tfrac{1}{6}\varepsilon^p; Q_\varepsilon^+)$$
$$+ P^\varepsilon(\tilde{\phi}_b - \tilde{x}_b > \tfrac{1}{6}\varepsilon^p; Q_\varepsilon^+).$$

The first term on the right side is estimated from (7.2) with $q = p$. The last two terms are estimated from Lemma 7.4 with $D = \{\hat{\phi}_b - \hat{x}_b > \varepsilon^p/6\} \cap Q_\varepsilon^+$ and $D = \{\tilde{\phi}_b - \tilde{x}_b > \varepsilon^p/6\} \cap Q_\varepsilon^+$. Thus, for small ε and suitable $k_1 > 0$

$$P^\varepsilon\left(\hat{x}_b - \tilde{x}_b > \frac{1}{2}\varepsilon^p; Q_\varepsilon^+\right) \leqq \exp\left(-\frac{k_1}{\sqrt{\varepsilon}}\right).$$

Similarly,

$$P^\varepsilon\left(\hat{x}_b - \tilde{x}_b < -\frac{1}{2}\varepsilon^p; Q_\varepsilon^+\right) \leqq \exp\left(-\frac{k_1}{\sqrt{\varepsilon}}\right).$$

Finally, using (7.3),

$$P^\varepsilon\left(|\tilde{x}_t - m_t^+| > \frac{1}{2}\varepsilon^p; B^+ \cap Q_\varepsilon^+\right) = P^+\left(|\tilde{x}_t - m_t^+|^q > \frac{1}{2}\varepsilon^{pq}; B^+ \cap Q_\varepsilon^+\right)$$
$$\leqq \frac{2N_q}{\varepsilon^{pq}} \cdot \varepsilon^q = 2N_q\varepsilon^{(1-p)q} < \frac{\varepsilon^l}{2}$$

for small ε, provided q is chosen large enough that $(1-p)q > l$. By combining these inequalities and recalling (5.14) we obtain Theorem 7.5.

8. Extensions of results. Let us outline how the quadratic variations and likelihood ratio tests can be modified to deal with more complicated situations. To begin with, let us again consider state x_t and observation y_t of dimension one. Afterward we outline an extension to state x_t and observation y_t of the same dimension $n > 1$.

Let us again assume (A1) of § 2. Instead of (A2) and (A3) we assume:

(A2′) h has a finite number of critical points x_i^*, $i = 1, \cdots, m$ with $x_1^* < \cdots < x_m^*$. Moreover, $|h(x)| \to \infty$ as $|x| \to \infty$.

To state the next assumption (A3′) let us make the convention that $x_0^* = \infty$, $x_{m+1}^* = +\infty$. Let

$$O_i = (x_{i-1}^*, x_i^*), \qquad i = 1, \cdots, m+1$$
$$N = \{x_1^*, \cdots, x_m^*\}.$$

(A3′) There exist $\gamma > 0$, $c > 0$ and for every $i < j$ such that $h(O_i) \cap h(O_j) \neq \varnothing$ a compact interval Δ_{ij} with the following properties:

(i) $\Delta_{ij} \subset h(O_i) \cap h(O_j)$;

(ii) $\gamma \leqq \operatorname{dist}(h(N), \Delta_{ij})$;

(iii) $x_i \in O_i$, $x_j \in O_j$, $h(x_i) = h(x_j) = y \in \Delta_{ij}$

imply

$$|[g(x_i)h'(x_i)]^2-[g(x_j)h'(x_j)]^2|>c. \tag{8.1}$$

In previous sections we considered $m=1$, $x_1^*=0$, $h(0)=0$, $\gamma=y_1$, $\Delta_{12}=[y_1, y_2]$.

Let us first describe a test like Lemma 4.1, which will imply that, on a time interval $a\leqq t\leqq b$, x_t remains outside a neighborhood of the set of critical points with the probability very nearly one. We choose $\delta>0$ such that

$$\text{dist}\,[h(x), h(N)]\geqq\gamma \quad \text{implies dist}\,(x, N)>2\delta. \tag{8.2}$$

As in § 4 we introduce "cutoffs" r and y_3, such that $\Delta_{ij}\subset[-y_3, y_3]$ for all $i<j$ and $y_3<|h(x)|$ whenever $|x|\geqq r$. Without loss of generality we may assume that $h'(x)<0$ on O_1.

Let

$$\begin{aligned}\Gamma_1&=(-r, x_1^*-\delta),\\ \Gamma_i&=[x_{i-1}^*+\delta, x_i^*-\delta], \qquad i=2,\cdots, m,\\ \Gamma_{m+1}&=[x_m^*+\delta, r].\end{aligned}$$

We choose h_i, $i=1,\cdots, m+1$, such that h_i'' is bounded and continuous, and

$$h_i(x)=h(x), \quad x\in\Gamma_i, \qquad 0<c_1\leqq|h_i'(x)|\leqq c_2. \tag{8.3}$$

We also define m_t^i for $t\geqq 0$, $i=1,\cdots, m+1$, by

$$dm_t^i=f(m_t^i)\,dt+(-1)^i g(m_t^i)\left(dy_t-\frac{1}{\varepsilon}h_i(m_t^i)\,dt\right) \tag{8.4}$$

with initial data $m_0^i=E^\varepsilon(x_0)$. As in (4.5) we define

$$\begin{aligned}&B_i=\{x_t\in\Gamma_i \text{ for } a\leqq t\leqq b\},\\ &C_\varepsilon^0=\{|h_I(m_t^I)|\leqq y_3, \text{dist}\,(h_I(m_t^I), h(N))\geqq\delta \text{ for } a\leqq t\leqq b\}\end{aligned} \tag{8.5}$$

for some particular I (e.g., $I=m+1$). The analogue of (4.6) is now

$$P^\varepsilon\left[\left(\bigcup_{i=1}^{m+1}B_i\right)^c\cap C_\varepsilon^0\right]\leqq\exp\left(-\frac{K}{\varepsilon}\right), \qquad 0<\varepsilon<\varepsilon_0. \tag{8.6}$$

If $h_I(m_t^I)\in\Delta_{ij}$ for $a\leqq t\leqq b$, then we need a test to discriminate between B_i and B_j. An analogue of Lemma 5.3 is as follows. Suppose that in (8.1) we have, as in (5.10a),

$$(gh')^2(x_i)<(gh')^2(x_j)-c. \tag{8.7}$$

If the opposite inequality holds, the discussion is similar. Let us set

$$\rho_s^-=(gh_i')(m_s^i), \qquad \rho_s^+=(gh_j')(m_s^j)$$

and define $C_\varepsilon=C_\varepsilon^{ij}=\{h_I(m_t^I)\in\Delta_{ij} \text{ for } a\leqq t\leqq b\}$. We again define $C_\varepsilon^+=C_\varepsilon^{+ij}$, $C_\varepsilon^-=C_\varepsilon^{-ij}$ by (5.11). Then, for small ε,

$$P^\varepsilon(B_i\cap C_\varepsilon^+)\leqq\exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad P^\varepsilon(B_j\cap C_\varepsilon^-)\leqq\exp\left(-\frac{K}{\sqrt{\varepsilon}}\right). \tag{8.8}$$

We then define $Q_\pm^\varepsilon=Q_\pm^{\varepsilon ij}$ as in (5.12) and obtain, as in (5.12), for small ε

$$P^\varepsilon(B_i|Q_\varepsilon^+)\leqq\exp\left(-\frac{K}{\sqrt{\varepsilon}}\right), \qquad P^\varepsilon(B_j|Q_\varepsilon^-)\leqq\exp\left(-\frac{K}{\sqrt{\varepsilon}}\right). \tag{8.9}$$

Similarly, given $i<j$ we define $L_\varepsilon = L_\varepsilon^{ij}$ as in (6.1) by

$$L_\varepsilon^{ij} = \int_d^b (\hat{h}_s^j - \hat{h}_s^i)\, dy_s - \frac{1}{2\varepsilon}\int_d^b [(\hat{h}_s^j)^2 - (\hat{h}_s^i)^2]\, ds$$

where $\hat{h}_s^i = h_i(m_s^i)$. If we again assume (8.7) then the analogue of Theorem 6.6 states that, for small ε and suitable K_m, $m = 1, 2, \cdots$,

$$P^\varepsilon(B_i \mid R_\varepsilon^+) \leqq K_m \varepsilon^m, \qquad P^\varepsilon(B_j \mid R_\varepsilon^-) \leqq K_m \varepsilon^m. \tag{8.10}$$

Remark. There is also a "local" version of these tests in which only $j = i+1$ is considered. This version can be used to detect a crossing of the critical point x_i^* by x_t, from O_i to O_{i+1} or vice versa.

Extensions to dimension $n>1$. Now let $h : \Re^n \to \Re^n$. We merely indicate how the results for $n=1$ can be modified, without spelling out in precise detail the results. Let $Dh(x)$ denote the matrix of partial derivatives and $Jh(x) = \det Dh(x)$ its Jacobian determinant. In the analogue of the (A1), the matrix $g(x)$ is assumed to have a bounded inverse $g^{-1}(x)$. Instead of (A2′) we suppose that there are disjoint open sets $O_i, \cdots, O_{m+1}$ such that

$$\Re^n = O_1 \cup \cdots \cup O_{m+1} \cup N,$$

$$N = \bigcup_{i=1}^{m+1} \partial O_i,$$

and the boundary of O_i consists of pieces of finitely many smooth $(n-1)$-dimensional manifolds. It is assumed that the restriction of h to each O_i is one-to-one with $Jh(x) \neq 0$ on O_i. Moreover, $|h(x)| \to \infty$ as $|x| \to \infty$.

Given $\gamma > 0$ we again choose $\delta > 0$ as in (8.2) and let

$$\Gamma_i = O_i \cap \{\text{dist}\,(x, N) \geqq \delta\} \cap \{|x| \leqq r\}$$

where r is a suitable "cutoff." We also choose h_i such that $h_i(x) = h(x)$ for $x \in \Gamma_i$ and that h_i satisfies the assumptions of Picard [13]. Principal among these is an analogue of (8.3), which asserts that h_i has an inverse h_i^{-1} with Dh_i^{-1} bounded. Let

$$T_i^2(x) = \text{trace}\,[(Dh_i)(gg^*)(Dh_i)^*](x). \tag{8.11}$$

Then $T_i^2(x) \geqq k_i > 0$. In (A3′), condition (8.1) is replaced by

$$|T_i^2(x_i) - T_j^2(x_j)| > c, \tag{8.12}$$

and Δ_{ij} is the closure of some bounded, open connected set (rather than an interval). As in Picard [13] the approximate filters m_t^i are defined by

$$dm_t^i = f(m_t^i)\, dt + [Dh_i(m_t^i)]^{-1} T_i(m_t^i)(dy_t - \varepsilon^{-1} h(m_t^i)\, dt) \tag{8.13}$$

with $m_0^i = E^\varepsilon(x_0)$. We then proceed as for $n=1$, with evident notational changes. For example, in the definition (5.6) we now have

$$Z_\varepsilon = \frac{1}{b-a} \sum_{j \text{ even}} |Y_{j+1} - Y_j|^2.$$

Acknowledgment. We thank J. Walsh for a helpful suggestion regarding the calculation above Lemma 5.1.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice Hall, Englewood Cliffs, NJ, 1979.
[2] R. J. Elliott, *Stochastic Calculus and Applications*, Springer-Verlag, Berlin, New York, 1982.

[3] W. H. FLEMING, D. JI, AND E. PARDOUX, *Piecewise Linear Filtering with Small Observation Noise*, Lecture Notes in Control and Information Science 3, Springer-Verlag, Berlin, New York, 1988, pp. 725-739.

[4] W. H. FLEMING, D. JI, P. SALAME, AND Q. ZHANG, *Discrete time piecewise linear filtering with small observation noise*, LCDS/CCS Report 88-27, Brown University, Providence, RI, 1988.

[5] U. HAUSSMAN AND E. PARDOUX, *A conditionally almost linear filtering problem with nongaussian initial condition*, Stochastics, 23 (1988), pp. 241-275.

[6] D. JI, *Asymptotic analysis of nonlinear filtering problems*, Ph.D. thesis, Brown University, Providence, RI, 1988.

[7] G. KALLIANPUR, *Stochastic Filtering Theory*, Springer-Verlag, Berlin, New York, 1980.

[8a] R. KATZUR, B. Z. BOBROVSKY, AND Z. SCHLUSS, *Asymptotic analysis of the optimal filtering problem for one-dimensional diffusions measured in a low noise channel, Part* I, SIAM J. Appl. Math., 44 (1984), pp. 591-604.

[8b] ———, *Asymptotic analysis of the optimal filtering problem for one-dimensional diffusions measured in a low noise channel, Part* II, SIAM J. Appl. Math., 44 (1984), pp. 1176-1191.

[9] A. J. KRENER, *The asymptotic approximation of nonlinear filters by linear filters*, in Proc. 7th MTNS Symposium, Stockholm, 1985.

[10] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[11] D. OCONE AND E. PARDOUX, *A Lie algebraic condition for nonexistence of finite dimensional computable filters*, in Proc. 2nd Trento Conference on Stochastic PDEs, 1988, Lecture Notes in Mathematics, Springer-Verlag, Berlin, New York, to appear.

[12] J. PICARD, *Nonlinear filtering of one-dimensional diffusions in the case of a high signal-to-noise ratio*, SIAM J. Appl. Math., 46 (1986), pp. 1098-1125.

[13] ———, *Filtrage de diffusions vectorielles faiblement bruitees*, Notes in Control and Information Science 84, Springer-Verlag, Berlin, New York, 1986.

[14] ———, *Methodes de perturbation pour les equations differentielles stochastiques et le filtrage nonlineaire*, these, Universite de Provence, Marseille, France, 1987.

[15] P. PRIOURET, *Processus de diffusion et equations differentielles stochastiques*, Lecture Notes in Mathematics 390, Springer-Verlag, Berlin, New York, 1973, pp. 38-114.

[16] S. S. SACHS, *Asymptotic analysis of linear filtering problems*, Ph.D. thesis, Case Western Reserve University, Cleveland, OH, 1980.

[17] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 5, pp. 1182–1219, September 1989

015

OPTIMAL FEEDBACK CONTROL OF INFINITE-DIMENSIONAL PARABOLIC EVOLUTION SYSTEMS: APPROXIMATION TECHNIQUES *

H.T. BANKS† AND C. WANG‡

This paper is dedicated to the memory of E.J. McShane.

Abstract. This paper presents a general approximation framework for the computation of optimal feedback controls in linear quadratic regulator problems for nonautonomous parabolic distributed parameter systems. This is done in the context of a theoretical framework using general evolution systems in infinite-dimensional Hilbert spaces. The authors discuss conditions for preservation under approximation of stabilizability and detectability hypotheses on the infinite-dimensional system. The special case of periodic systems is also treated.

Key words. evolution equations, LQR problem, feedback control, approximation techniques

AMS(MOS) subject classifications. 49B22, 49B27, 65M60, 93C50, 93D15

1. Introduction.

In this paper we present a theoretical approximation framework for the computation of optimal feedback controls in linear quadratic regulator (LQR) problems governed by parabolic partial differential equations with time dependent coefficients. Our efforts were originally motivated by the desire to develop control strategies (distributed in nature) for insect dispersal models (see Chapter 1 of [BK2] and the references therein) which have been shown to involve time dependent coefficients.

The presentation below is somewhat in the spirit of that for autonomous parabolic systems in [BK1] and [LT] in that we attempt to develop a convergence theory in which uniform stabilizability of the original system is preserved under approximation. It differs substantially from [BK1] and [LT] since we do not directly use sectorial properties of the operators and resolvent and spectral set arguments to establish preservation of stabilizability and detectability. (Indeed, the time dependent nature of our system prevents this.) Nor do we use the Trotter–Kato theorem (which is not well suited for use with nonautonomous control systems) in our convergence arguments.

In §2 we summarize previous results for abstract LQR problems on infinite time intervals and formulate these in a form readily used in our subsequent discussions. This formulation is based on the abstract frameworks found in [CP], [G], [BK1], and [Da], [DI1], [DI2]; we rely heavily on the ideas of Da Prato and Ichikawa that guarantee uniqueness of solutions of the associated Riccati integral equations under certain stabilizability and detectability assumptions.

An approximation framework for abstract evolution systems in the spirit of [G] and [BK1] is given in §3; convergence of the approximate Riccati operators (and, of course, the corresponding controls and trajectories) is established under uniform

* Received by the editors December 28, 1988; accepted for publication January 16, 1989. Research for this paper was supported in part under grants NSF MCS8504316, NASA NAG-1-517, AFOSR-84-0398, and AFOSR-F49620-86-C-0111.

† Center for Control Sciences, Division of Applied Mathematics, Brown University, Providence, Rhode Island, 02912. Part of this research was carried out while the first author was a visiting scientist at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, which is operated under NASA contract NAS1-18107.

‡ Department of Mathematics, University of Southern California, University Park, Los Angeles, California 90089-1113

stabilizability and detectability hypotheses on the approximate evolution control systems.

Our major contributions are given in §4, along with the presentations in §§5 and 6, where we show how the hypotheses of §4 can be verified for rather wide classes of problems of interest. In §4 we focus attention on parabolic systems described by time dependent sesquilinear forms (in the spirit of the autonomous system frameworks in [BK1], [BI1], [BI2]) and associated evolution equations. We make substantial use of the results of Tanabe [T] to formulate our problems in a weak (V^*) sense. Our fundamental convergence results (Theorem 4.4) for the uncontrolled systems rely on a sesquilinear or variational formulation of the systems, strong V-ellipticity of the parabolic evolution systems, approximation properties for the spaces approximating the state space, and the Gronwall inequality. (Certain aspects of this approach can be relaxed to allow us to treat weakly damped hyperbolic systems—see [BKS], [BKW].) We are then able to reduce convergence questions for the controlled systems (e.g., convergence of Riccati variables, optimal controls, and feedback evolution systems) to conditions of uniform stabilizability and uniform detectability of the approximate systems (Theorem 4.5).

We show in §5 that we can obtain these uniform stabilizability/detectability conditions by preservation under approximation of dissipative inequalities for certain classes of evolution control systems. Sufficient conditions that are readily checked in many examples are given and several special cases are noted.

An alternative approach is presented in §6 where we restrict our considerations to parabolic systems for which the domain V of the generator of the evolution system embeds compactly in the state space H. In this case, it is shown that stabilizability/detectability of the original system is preserved under approximation.

Finally, in §7 we give an example of a class of parabolic partial differential equation control problems for which all the hypotheses of our theoretical framework can be easily verified.

We have used the ideas presented in this paper to develop and test computational packages for solving nonautonomous parabolic control problems of the type discussed in §7. However, since our presentation here is already quite long and since a presentation of our detailed numerical findings would entail lengthy discussions, we will not discuss the numerical examples. A separate manuscript is under preparation; the interested reader can also consult [W].

We believe that the present paper offers new results for time dependent infinite-dimensional control systems. Moreover, our arguments are such that we offer an attractive alternative approach to those found in [G], [BK1], and [LT] even in the case of autonomous parabolic systems.

2. The abstract linear quadratic regulator problem on an infinite time interval. In this section we formulate a linear quadratic regulator problem for evolution system dynamics in a Hilbert space and present a collection of functional analytic and control theoretic results related to such problems. The results we give in this section are known even though in some cases we have modified the statements to present the results in a form most suited to our purposes. The reader can easily refer to the cited literature for proofs. In particular, we use freely results found in [CP] and [G] and rely heavily on recent results of Da Prato and Ichikawa [Da], [DI1], and [DI2].

We first recall results for evolutionary systems. Let H be a Hilbert space with inner product $< \cdot, \cdot >$. Let $\triangle(t_0, t_f) = \{(t,s) \mid t_0 \leq s \leq t \leq t_f\}$, $\triangle_\infty(t_0) = \{(t,s) \mid t_0 \leq s \leq t < \infty\}$ and $L(H)$ be the Banach algebra of bounded linear operators

on H. We use $B_\infty([t_0,t_f];L(H))$ to denote the set of strongly measurable operator valued functions that are bounded on $[t_0,t_f]$. We recall that $T(\cdot,\cdot):\triangle(t_0,t_f)\mapsto L(H)$ is called an *evolution operator* if T satisfies the following conditions: (i) $T(t,s)=T(t,r)T(r,s)$, for $t_0\le s\le r\le t\le t_f$; (ii) $T(t,t)=I$, for $t\in[t_0,t_f]$; and (iii) $T(t,s)$ is strongly continuous in s on $[t_0,t]$ and strongly continuous in t on $[s,t_f]$. We say that an evolution operator has *exponential growth* if there exists $M_1\ge 1,\omega>0$ such that $\|T(t,s)x\|\le M_1e^{\omega(t-s)}\|x\|$, for $(t,s)\in\triangle_\infty(t_0),x\in H$. An evolution operator is said to be *uniformly exponentially stable* if there exists $M\ge 1$ and $\alpha>0$ such that $\|T(t,s)x\|\le Me^{-\alpha(t-s)}\|x\|$, for $(t,s)\in\triangle_\infty(t_0),x\in H$. We have the following fundamental results of Datko which will be crucial to our presentation.

LEMMA 2.1. [Dt]. *Consider an evolution operator $T(\cdot,\cdot)$ with exponential growth. Then $T(\cdot,\cdot)$ is uniformly exponentially stable if and only if there exists an M_2 such that*

$$\int_s^\infty \|T(t,s)x\|^2dt\le M_2\|x\|^2,\quad \textit{for } s\ge t_0, x\in H.$$

Furthermore, we can find constants $M_3\ge 1,\alpha>0$ depending only on M_1,M_2, and ω for which the following estimate holds:

$$\|T(t,s)\|\le M_3e^{-\alpha(t-s)},\quad \textit{for } (t,s)\in\triangle_\infty(t_0).$$

The original statement and proof of this theorem are due to Datko. We have modified slightly (see Appendix A of [W]) the original proof in [Dt] to point out the relationship between the constants M_3,α and M_1,M_2, and ω. This will be essential for our subsequent use with approximation systems. In [W] it is shown that the constants M_3 and α can be chosen as:

$$M_3=2M_1e^{4M_2{M_1}^2\omega(2\omega M_2+1)},\quad \alpha=\frac{\log 2}{4M_2{M_1}^2(2\omega M_2+1)}.$$

In our discussions of control systems, perturbations of evolution operators (see [CP]) will play an important role. Let $t_f<\infty$. Consider a uniformly bounded evolution operator $T(\cdot,\cdot)$ and $C(\cdot)\in B_\infty([t_0,t_f];L(H))$. Then the integral equation for $S(t,s)\in L(H)$ given by

$$S(t,s)x=T(t,s)x+\int_s^t T(t,\eta)C(\eta)S(\eta,s)xd\eta,\quad \text{for } x\in H,$$

has a unique solution $S(\cdot,\cdot)$ in the class of strongly continuous operator valued functions. Moreover, $S(\cdot,\cdot)$ is an evolution operator and is called the *perturbed evolution operator* corresponding to the perturbation of $T(\cdot,\cdot)$ by $C(\cdot)$. In addition, $S(\cdot,\cdot)$ is also the unique solution of

$$S(t,s)x=T(t,s)x+\int_s^t S(t,\eta)C(\eta)T(\eta,s)xd\eta,\quad \text{for } x\in H.$$

We turn next to our formulation of the regulator problem for an evolution system. We let H,U be real Hilbert spaces with inner products $<\cdot,\cdot>_H$, $<\cdot,\cdot>_U$; H,U are the state space and the control space, respectively. Consider an evolution operator

$T(\cdot,\cdot)$ defined on $\triangle_\infty(t_0)$. For any $u \in L^2([t_0,\infty);U)$, the control system trajectories are defined by:

$$x(t) = T(t,s)x(s) + \int_s^t T(t,\tau)B(\tau)u(\tau)d\tau, \quad \text{for } (t,s) \in \triangle_\infty(t_0). \tag{2.1}$$

The cost functional is given by

$$J_\infty(u;t_0,x_0) = \int_{t_0}^\infty \{< W(t)x(t),x(t) >_H + < R(t)u(t),u(t) >_U\}\, dt, \tag{2.2}$$

where $x(\cdot)$ is the trajectory corresponding to u with $x(t_0) = x_0$. For each given t_0, x_0, the optimal control problem is to find a control u^* which minimizes (2.2) over all $u \in L^2([t_0,\infty);U)$.

We can consider (2.2) as the limit as $t_k \to \infty$ of

$$\begin{aligned} J(u;t_0,t_k,x_0) =& < Gx(t_k),x(t_k) >_H \\ &+ \int_{t_0}^{t_k} \{< W(t)x(t),x(t) >_H + < R(t)u(t),u(t) >_U\}\, dt, \end{aligned} \tag{2.3}$$

with $G = 0$. Here we shall summarize existence results for optimal controls in the infinite time interval, existence and uniqueness of the solutions of the Riccati integral equation on an infinite time interval, and stability of the feedback system.

We make the following *standing assumptions* for all subsequent discussions of (2.1), (2.2): (i) The evolution operator $T(\cdot,\cdot)$ has exponential growth. (Thus, in particular, $T(t,s)$ is uniformly bounded for s,t in any bounded sub-interval of $[t_0,\infty)$); (ii) The strongly measurable operator valued function $B(\cdot) : [t_0,\infty) \mapsto L(U,H)$ is uniformly bounded in $[t_0,\infty)$, i.e., there exists M_B such that $\|B(t)\|_{L(U,H)} \le M_B$ for all $t \in [t_0,\infty)$; (iii) The strongly measurable operator valued function $W(\cdot) : [t_0,\infty) \mapsto L(H)$ is uniformly bounded in the interval $[t_0,\infty)$, and $W(t)$ is nonnegative definite self-adjoint for all $t \in [t_0,\infty)$; (iv) The strongly measurable operator valued function $R(\cdot) : [t_0,\infty) \mapsto L(U)$ is uniformly bounded in the interval $[t_0,\infty)$, and $R(t)$ is positive definite self-adjoint for all $t \in [t_0,\infty)$. Furthermore, there exists a constant $r > 0$ such that $< R(t)u,u >_U \ge r\|u\|_U^2$, for all $u \in U$ and $t \ge t_0$.

Under these assumptions, we consider the linear quadratic control problem in the interval $[t_0,t_k]$ for $t_k < \infty$. That is, we consider the cost functional (2.3) with our system (2.1). Then for any bounded self-adjoint nonnegative definite linear operator G, it is well known that for each given $x_0 \in H$, there exists a unique control u such that

$$J(u;t_0,t_k,x_0) \le \min_{v \in L^2([t_0,t_k];U)} J(v;t_0,t_k,x_0).$$

This control u can be written in a feedback form $u(t) = -R^{-1}(t)B^*(t)Q(t)x(t)$, for $t \in [t_0,t_k]$, where $x(\cdot)$ is the corresponding trajectory and $Q(\cdot) : [t_0,t_f] \mapsto L(H)$, is the unique self-adjoint solution of the Riccati integral equation (RIE)

$$\begin{aligned} Q(t)x =& T^*(t_k,t)GT(t_k,t)x + \int_t^{t_k} T^*(\eta,t)W(\eta)T(\eta,t)x\,d\eta \\ &- \int_t^{t_k} T^*(\eta,t)Q(\eta)B(\eta)R^{-1}(\eta)B^*(\eta)Q(\eta)T(\eta,t)x\,d\eta \end{aligned} \tag{2.4}$$

for all $t \in [t_0, t_k]$ and $x \in H$.

We note that in the case $G = 0$, the above equation reduces to

$$Q(t)x = \int_t^{t_k} T^*(\eta, t)[W(\eta) - Q(\eta)B(\eta)R^{-1}(\eta)B^*(\eta)Q(\eta)]T(\eta, t)x d\eta, \tag{2.5}$$

for all $t \in [t_0, t_k]$ and $x \in H$. We also note that (2.4) is equivalent to

$$\begin{aligned} Q(s)x &= T^*(t,s)Q(t)T(t,s)x + \int_s^t T^*(\eta,s)W(\eta)T(\eta,s)x d\eta \\ &\quad - \int_s^t T^*(\eta,s)Q(\eta)B(\eta)R^{-1}(\eta)B^*(\eta)Q(\eta)T(\eta,s)x d\eta \\ Q(t_k)x &= Gx \end{aligned} \tag{2.6}$$

for all $t_0 \le s \le t \le t_k$ and $x \in H$. Solutions of this latter equation have a representation that is often used in control theoretic arguments. Consider any $u(\cdot) \in L^2([t_0, t_k]; U)$, and for $x \in H$, define a H-valued function $y(\cdot)$ by

$$y(t) = T(t,s)x + \int_s^t T(t,\tau)B(\tau)u(\tau)d\tau, \quad \text{for } t \in [s, t_k].$$

If $Q(\cdot)$ is a self-adjoint solution of (2.6), then

$$\begin{aligned} < Q(s)x, x >_H &= < Gy(t_k), y(t_k) >_H \\ &\quad + \int_s^{t_k} \{< W(t)y(t), y(t) >_H + < R(t)u(t), u(t) >_U\} dt \\ &\quad - \int_s^{t_k} < R(t)z(t), z(t) >_U dt, \end{aligned} \tag{2.7}$$

where $z(t) = u(t) + R^{-1}(t)B^*(t)Q(t)y(t)$. This can be used to show that (2.6) has a unique self-adjoint solution.

Before continuing our discussion, let us introduce additional notation. Let

$$\Sigma^+ = \left\{ E \mid E \in L(H), E \text{ self-adjoint, nonnegative definite.} \right\}$$
$$\mathcal{C}_s([t_0, t_k]; \Sigma^+) = \left\{ K : [t_0, t_k] \mapsto \Sigma^+ \mid K \text{ strongly continuous.} \right\}$$

By the uniqueness of the solution of (2.6), we can define a mapping $\Lambda : \Sigma^+ \mapsto \mathcal{C}_s([t_0, t_k]; \Sigma^+)$ as follows: for each $G \in \Sigma^+$, ΛG is the unique nonnegative definite self-adjoint solution of (2.6). Under our general assumptions, it is easily seen that for fixed G the map Λ depends only on t_k; if we consider the linear quadratic regulator problem on two bounded intervals $[t_0, t_1]$ and $[t_0, t_2]$, we will use Λ_1, Λ_2 to denote the maps associated with each interval, respectively.

Now consider a increasing sequence $\{t_k\}_{k=1}^{\infty}$, with $t_k < \infty$. The map Λ_k associates with each finite interval problem the Riccati equation on $[t_0, t_k]$. Let $G = 0$ and $Q_k(\cdot) = \Lambda_k G$. For simplicity, consider a bounded interval $[a, b] \subset [t_0, t_1]$, and for each $t \in [a, b], x \in H$, we assume that there exists a constant $M(t, x)$ such that for all k

$$< Q_k(t)x, x >_H \quad \le \quad M(t, x). \tag{2.8}$$

The following theorem (see [Da], [G] establishes the connection between control problems on a finite time interval and problems on an infinite time interval.

THEOREM 2.1. *Under the standing assumptions and* (2.8), *we can conclude the following:*

(i) *For each* $t \in [a,b]$, *there exists a unique operator* $Q(t) \in \Sigma^+$ *such that* $Q_k(t) \to Q(t)$ *strongly and the convergence is uniform in* $[a,b]$. *Therefore,* $Q(\cdot)$ *is strongly continuous, so uniformly bounded in* $[a,b]$.

(ii) *As a consequence of (i), we can define the perturbed evolution systems* $S_k(\cdot,\cdot)$, $S(\cdot,\cdot)$ *corresponding to the perturbation of* $T(\cdot,\cdot)$ *by* $-BR^{-1}B^*Q_k$ *and* $-BR^{-1}B^*Q$, *respectively. We have* $S_k(t,s)x \to S(t,s)x$, *for all* $x \in H$, *and* $a \le s \le t \le b$. *Furthermore, the convergence is uniform in* t *for* $t \in [s,b]$. *If* $T(\cdot,\cdot)$ *is jointly strongly continuous, then the convergence is uniform for all* $a \le s \le t \le b$.

The only assumption on the sequence $\{t_k\}$ is that t_k increase as a function of k. In particular, the above theorem is valid when $t_k \to \infty$, as $k \to \infty$. Paralleling the usual approach to finite-dimensional regulator problems, we can use these results to establish results for the control problem on an infinite time interval. To that end, consider a sequence $\{t_k\}_{k=1}^{\infty}$ with $t_k \to \infty$ as $k \to \infty$. Let $Q_k(\cdot), S_k(\cdot,\cdot)$ be defined as above. If for each $t \ge t_0$ we can find a constant $M(t)$ such that $< Q_k(t)x, x >_H \le M(t)\|x\|^2$, then by Theorem 2.1, we have $Q(\cdot), S(\cdot,\cdot)$ defined on $[t_0,\infty)$. Furthermore, for any $(t,s) \in \Delta_\infty(t_0)$, Q satisfies

$$
\begin{aligned}
(2.9) \qquad Q(s)x &= T^*(t,s)Q(t)T(t,s)x + \int_s^t T^*(\eta,s)W(\eta)T(\eta,s)x\,d\eta \\
&\quad - \int_s^t T^*(\eta,s)Q(\eta)B(\eta)R^{-1}(\eta)B^*(\eta)Q(\eta)T(\eta,s)x\,d\eta.
\end{aligned}
$$

Equation (2.9) is called the Riccati integral equation (RIE) for the infinite time interval. We know from Theorem 2.1 that Q is strongly continuous and uniformly bounded in any bounded interval, but Q is not necessarily uniformly bounded in the entire interval $[t_0,\infty)$. If Q is not uniformly bounded, that implies the minimal cost for some initial state x will tend to infinity as t_k tends to infinity; that is, there is no control yielding finite cost for the infinite time interval problem. Let us state a condition which prohibits this situation.

DEFINITION 2.1. (W-stabilizability). We say that (2.1), (2.2) is *W-stabilizable* if there exists a constant M such that for any $s \ge t_0$ and $x \in H$, we can find a control $u \in L^2([t_0,\infty);U)$ satisfying

$$
(2.10) \qquad J_\infty(u;s,x) \le M\|x\|^2.
$$

One can then prove (see [DI1, Thm. 3.1]) that $Q = \lim Q_k$ is a uniformly bounded solution of the Riccati integral equation (2.9) in $[t_0,\infty)$ if and only if (2.1), (2.2) is W-stabilizable. In this case we have $Q(t) \le M \cdot I$ for $t \in [t_0,\infty)$. Furthermore, if $\hat{Q}$ is any other bounded self-adjoint solution of (2.9), we have that $Q(t) \le \hat{Q}(t)$ for $t \in [t_0,\infty)$. It follows that using any sequence $\{t_k\}$, with $t_k \to \infty$, in the above limiting procedure yields the same solution Q to (2.9), which we shall refer to as the

minimal bounded nonnegative self-adjoint solution of the Riccati integral equation on $[t_0, \infty)$ and denote by $Q_{\min}$.

We note that if the system (2.1), (2.2) is W-stabilizable, then for any $s \geq t_0$ and $x \in H$, the unique optimal control for the infinite time interval problem is given by $u(t) = -R^{-1}(t)B^*(t)Q(t)S(t,s)x$.

Next we consider a uniformly bounded solution $\hat{Q}$ of (2.9) and let $\hat{S}$ be the evolution operator corresponding to the perturbation of T by $-BR^{-1}B^*\hat{Q}$. We say that $\hat{Q}$ is a *stability solution* of (2.9) if $\hat{S}(t,s)x \to 0$ as $t \to \infty$ for all $s \geq t_0$, $x \in H$.

It is shown in [DI1] that there is at most one stability solution of (2.9). Moreover, if $\hat{Q}$ is a stability solution satisfying $\hat{Q}(t) \leq M \cdot I$ and Q_k is the solution on $[t_0, t_k]$ with $Q_k(t_k) = M \cdot I$, then $\hat{Q}(t) \leq Q_k(t)$ for $t \in [t_0, t_k]$ and $Q_k(t)x \to \hat{Q}(t)x$ as $k \to \infty$ for each $x \in H$. In addition, if Q is any uniformly bounded solution, then $Q(t) \leq \hat{Q}(t)$, $t \in [t_0, \infty)$; that is, any stability solution is the *maximal* (uniformly bounded) solution. Finally, if the system (2.1), (2.2) is W-stabilizable and if the minimal solution $Q_{\min}$ of (2.9) is a stability solution, then it is the unique uniformly bounded solution of the RIE (2.9).

From the above remarks, it is clear that it is desirable to have conditions that guarantee a solution of the RIE be a stability solution. One such condition is a detectability condition which plays a role in infinite dimensional systems that is analogous to its role in finite dimensional systems (see [R], [DI1]).

DEFINITION 2.2. (W-detectability). Let $V(t) = \sqrt{W(t)}$. We say that the system (2.1), (2.2) is *W-detectable* if there exists a uniformly bounded function $K(\cdot)$ with $K(t) \in L(H)$ such that the evolution operator T_{KV} corresponding to the perturbation of T by KV is uniformly exponentially stable.

We then have the following result.

THEOREM 2.2. *Suppose that the system* (2.1), (2.2) *is W-stabilizable and W-detectable. Then the minimal solution $Q_{\min}$ of the* RIE *is the unique uniformly bounded solution of* (2.9) *and the evolution operator S defined by perturbation of T by $-BR^{-1}B^*Q_{\min}$ is uniformly exponentially stable. In fact,*

$$\|S(t,s)\| \leq Me^{-\alpha(t-s)}, \quad (t,s) \in \triangle_\infty(t_0),$$

where the constants M and α depend only on the bounds for $B, K, R^{-1}, Q_{\min}$, and M_{KV}, β in the bound $\|T_{KV}(t,s)\| \leq M_{KV} \exp\{-\beta(t-s)\}$.

The first part of this theorem follows from [DI1, Prop. 3.3]. That the constants M and α depend only on the bounds indicated follows from use of the modified Datko lemma, Lemma 2.1 above. As we shall see in the next section on approximation, this dependence (or lack thereof) will allow us to infer a uniform exponential stability of the approximate feedback control systems whenever we have a uniform W-detectability condition satisfied by the approximate systems.

To conclude this section, we recall that an evolution operator is said to be θ-periodic if for any $(t,s) \in \triangle_\infty(t_0)$, we have $T(t+\theta, s+\theta)x = T(t,s)x$, for all $x \in H$. We note that any θ-periodic evolution operator satisfies the exponential growth assumption that is part of our standing assumptions in this paper. It is also easily argued (e.g., see [W]) that if the linear quadratic regulator problem is θ-periodic (i.e., B, W, R, and T of (2.1), (2.2) are θ-periodic), then the minimal solution and the stability solution of the RIE are θ-periodic. Of course, we cannot argue that every

uniformly bounded solution of the RIE is periodic under a periodicity assumption on the problem.

We turn next to approximation results for the abstract linear regulator problem on an infinite time interval.

3. Approximation of linear quadratic regulator problems on an infinite time interval. Let H^N and U^N be families of finite-dimensional subspaces of the original state space and control space H, U, respectively. For each N an approximate control system is described by

$$x^N(t) = T^N(t,s)x^N(s) + \int_s^t T^N(t,\eta)B^N(\eta)u^N(\eta)d\eta, \quad \text{for } (t,s) \in \triangle_\infty(t_0), \tag{3.1}$$

where $T^N(\cdot,\cdot) : \triangle_\infty(t_0) \mapsto L(H^N)$ is an evolution operator, and $B^N(\cdot) : U^N \mapsto H^N$. The cost functional is given by

$$\begin{aligned} J_\infty^N(u^N; t_0, x_0^N) = &\int_{t_0}^\infty < W^N(t)x^N(t), x^N(t) >_H \, dt \\ &+ \int_{t_0}^\infty < R^N(t)u^N(t), u^N(t) >_U \, dt \end{aligned} \tag{3.2}$$

where $x^N(\cdot)$ satisfies (3.1) and $x^N(t_0) = x_0^N$. Suppose that each of the approximate systems and cost functionals satisfies the standing assumptions for (2.1), (2.2) given above and that each is W-stabilizable. Then we can guarantee existence of $Q^N(\cdot)$, the minimal uniformly bounded solution of the associated Riccati integral equation on the infinite time interval $[t_0, \infty)$. Let $S^N(\cdot,\cdot)$ be the perturbed evolution operator corresponding to the perturbation of T^N by $-B^N(R^N)^{-1}B^{*N}Q^N$. In this section, we present results on the strong convergence of Q^N, S^N.

We need to make some basic assumptions on the approximate systems. Let $\{H^N\}_{N=1}^\infty$, $\{U^N\}_{N=1}^\infty$ be subspaces of H, U, respectively, and P_H^N, P_U^N be projection operators which are assumed to satisfy $\|P_H^N x - x\|_H \to 0$, $\|P_U^N u - u\|_U \to 0$, as $N \to \infty$, for all $x \in H, u \in U$.

We note that the usual orthogonal projections of H and U onto H^N, U^N respectively, satisfy these assumptions if H^N, U^N approximate H and U in an appropriate sense. (We shall specify approximation systems that satisfy these conditions in subsequent sections.) We make the further assumptions on our approximate systems.

Hypothesis 3.1. (*Uniform boundedness*).

(i) There exist constants $M \geq 1$ and $\omega > 0$ such that

$$\|T(t,s)\|_{L(H)} \leq Me^{\omega(t-s)}, \quad \|T^N(t,s)\|_{L(H^N)} \leq Me^{\omega(t-s)}$$

hold for all N and $(t,s) \in \triangle_\infty(t_0)$;

(ii) There exists a constant K_B such that

$$\|B(t)\|_{L(U,H)} \leq K_B, \quad \|B^N(t)\|_{L(U^N,H^N)} \leq K_B$$

for all N and $t \in [t_0, \infty)$;

(iii) There exists a constant K_W such that

$$\|W(t)\|_{L(H)} \leq K_W, \quad \|W^N(t)\|_{L(H^N)} \leq K_W$$

for all N and $t \in [t_0, \infty)$. Furthermore, $W(t), W^N(t)$ are nonnegative definite self-adjoint for all $t \in [t_0, \infty)$.

(iv) There exists a constant K_R such that

$$\|R(t)\|_{L(U)} \leq K_R, \quad \|R^N(t)\|_{L(U^N)} \leq K_R$$

for all N and $t \in [t_0, \infty)$. In addition, $R(t), R^N(t)$ are positive definite self-adjoint for all $t \in [t_0, \infty)$. There exists a constant $r > 0$ such that $R(t) \geq r \cdot I$, $R^N(t) \geq r \cdot I$, for all $t \in [t_0, \infty)$.

Hypothesis 3.2. (*Pointwise convergence*). The operators $T^N(t,s)$, $T^{*N}(t,s)$, $B^N(t)$, $B^{*N}(t)$, G^N, $W^N(t)$, $R^N(t)$ converge strongly to $T(t,s)$, $T^*(t,s)$, $B(t)$, $B^*(t)$, G, $W(t)$, $R(t)$ for any $t_0 \leq s \leq t < \infty$, where G, G^N are nonnegative self-adjoint operators in $L(H), L(H^N)$, respectively.

From arguments in [DI1] and [W], it is readily seen that W-stabilizability (i.e., condition (2.10)) is equivalent to the following: there exists a constant $M > 0$, and a uniformly bounded feedback operator $K(\cdot) : [t_0, \infty) \mapsto L(H, U)$ such that if $T_K(\cdot, \cdot)$ is the perturbed evolution operator corresponding to the perturbation of T by BK, then for any $s \geq t_0$, $x \in H$, the cost of the feedback control $u(t) = B(t)K(t)T_K(t,s)x$ satisfies $J_\infty(u; s, x) \leq M\|x\|^2$.

To guarantee the existence of uniformly bounded solutions of the Riccati integral equation on the infinite time interval for each of the approximate systems, we make a uniform W-stabilizability assumption.

Hypothesis 3.3. (*Uniform W-stabilizability*). There exists a constant $M > 0$ such that for all N, there exist uniformly bounded feedback operators $K^N(\cdot) : [t_0, \infty) \mapsto L(H^N, U^N)$ satisfying the following: for all $s \geq t_0$ and $x^N \in H^N$, the feedback control $u^N(t) = B^N(t)K^N(t)T^N_{K^N}(t,s)x^N$ has a cost satisfying

$$J^N_\infty(u^N; s, x^N) \leq M\|x^N\|^2_H.$$

Now consider $\{t_k\}_{k=1}^\infty$ with $t_k \to \infty$ as $k \to \infty$. For each N, let Σ^{+N} be the set of nonnegative self-adjoint linear operators in H^N; we define the map $\Lambda^N_k : \Sigma^{+N} \mapsto \mathcal{C}_s([t_0, t_k]; \Sigma^{+N})$ via the finite dimensional Riccati integral equation on $[t_0, t_k]$ as before. Let $G \in \Sigma^+$, and $G^N \in \Sigma^{+N}$. Define $Q^N_k(\cdot) = \Lambda^N_k G^N$, and $Q_k(\cdot) = \Lambda_k G$. Let the evolution operators S^N_k and S_k correspond to the perturbation of T^N, T by $-B^N(R^N)^{-1}B^{*N}Q^N_k$ and $-BR^{-1}B^*Q_k$, respectively. The theories of the approximation of linear quadratic control problems on a finite time interval (e.g., see [G], [BK1]) guarantee that under Hypotheses 3.1 and 3.2, for each k, $Q^N_k(t)$ and $S^N_k(t,s)$ converge strongly to $Q_k(t)$ and $S_k(t,s)$, respectively, as $N \to \infty$ for every $t_0 \leq s \leq t \leq t_k$. Furthermore the convergence is uniform in the interval $[t_0, t_k]$, if we replace Hypothesis 3.2 by the following assumptions.

Hypothesis 3.4. (*Continuity and uniform convergence*). The operator valued functions $B(t)$, $B^*(t)$, $W(t)$, $R(t)$ are strongly piecewise continuous in t (with only a finite number of discontinuity points in any bounded interval); the evolution operators T, T^* are jointly strongly continuous. The convergences in Hypothesis 3.2 are uniform in t and (t,s) on any bounded interval.

From the theory of the linear quadratic control problem for the infinite time intervals, there are two cases where Q^N_k converges strongly as $k \to \infty$. In the first

case, let $Q^N_{\min}(t)$ be the minimal uniformly bounded solution of the Riccati integral equation in H^N on the infinite time interval $[t_0, \infty)$. If $G = G^N = 0$, then for any N we have $Q^N_k(t)x^N \to Q^N_{\min}(t)x^N$, as $k \to \infty$. Furthermore the convergence is uniform in t for t in any bounded interval $[t_0, t_f]$. In the second case, we assume the following conditions hold.

Hypothesis 3.5. Assume that there exists a stability solution $Q_s(\cdot)$ of the Riccati integral equation for the infinite time interval infinite-dimensional system and there exists a stability solution $Q^N_s(\cdot)$ of the Riccati integral equation on the infinite time interval for each approximate system. (Then the evolution operator S^N corresponding to the perturbation of T^N by $-B^N(R^N)^{-1}B^{*N}Q^N_s$ satisfies $S^N(t,s)x \to 0$, as $t \to \infty$, for $s \geq t_0, x \in H^N$.) Furthermore, assume that there exists a constant M such that for each N, $Q^N_s(t) \leq M \cdot I$ for all $t \geq t_0$. Also assume $Q_s(t) \leq M \cdot I$ for all $t \geq t_0$.

Assuming that Hypothesis 3.5 holds, we let $G = G^N = M \cdot I$ and Q^N_k, Q_k be the solutions of the RIE on $[t_0, t_k]$ satisfying $Q^N_k(t_k) = G^N$, $Q_k(t_k) = G$. Then from our results for stability solutions given in §2, we have $Q^N_k(t)x^N \to Q^N_s(t)x^N$, $Q_k(t)x \to Q_s(t)x$ as $k \to \infty$ for all $x^N \in H^N$ and $x \in H$.

We note that if $S^N(t,s)x \to 0$ uniformly in N, then we have $Q_k(s)$ and $Q^N_k(s)$ uniformly bounded for all k and N. To see this, we consider Q_k, Q_s as given above. Then we have, using the relationship in (2.7) where $y(t) = S(t,s)x$ and S is the evolution operator corresponding to the perturbation of T by $-BR^{-1}B^*Q_s$,

$$\begin{aligned} < (Q_k(s) - Q_s(s))x, x >_H &\leq < (Q_k(t_k) - Q_s(t_k))y(t_k), y(t_k) >_H \\ &\leq 2M\|y(t_k)\|^2. \end{aligned}$$

Since $y(t_k) \to 0$, it follows from the uniform boundedness principle that $Q_k(s)$ is uniformly bounded for all k. Repeating this argument with $Q^N_k(s), Q^N_s(s)$ and $y^N(t) = S^N(t,s)x$, we see that the uniform (in N) decay of S^N yields the claimed uniform boundedness for $Q^N_k(s)$.

In each of the two cases above, we have the following situation:

$$\begin{array}{ccc} Q^N_k(t)P^N_H x & \overset{k\to\infty}{\to} & Q^N_s(t)P^N_H x \\ N \to \infty \downarrow & & \downarrow ?N \to \infty \\ Q_k(t)x & \overset{k\to\infty}{\to} & Q_s(t)x. \end{array}$$

It is desirable in computations to work directly with Q^N_s and hence we seek results which will guarantee the convergence $Q^N_s \to Q_s$ of this diagram. To obtain such a result, we shall make use of a uniform decay rate for the S^N defined via Q^N_s.

THEOREM 3.1. *Assume that Hypotheses* 3.1–3.3, 3.5 *hold. Further, assume that for all* $s \geq t_0$, $x \in H$, *and* $\epsilon > 0$, *we can find* $\hat{t}$ *such that for all* $t \geq \hat{t}$, *we have* $\|S(t,s)x\| \leq \epsilon$ *and* $\|S^N(t,s)P^N_H x\| \leq \epsilon$ *for all* N. *Then* $Q^N_s(t)P^N_H x \to Q_s(t)x$ *for all* $t_0 \leq t < \infty$.

Proof. Let M be the bound for Q_s and Q^N_s that are the stability solutions of Hypothesis 3.5. Let Q^N_k and Q_k be the related RIE solutions on $[t_0, t_k]$ satisfying $Q^N_k(t_k) = M \cdot I$, $Q_k(t_k) = M \cdot I$. Then for $t \leq t_k$ we have

$$\begin{aligned} \|Q^N_s(t)P^N_H x - Q_s(t)x\|^2 &\leq \|(Q^N_s(t) - Q^N_k(t))P^N_H x\|^2 + \|Q^N_k(t)P^N_H x - Q_k(t)x\|^2 \\ &\quad + \|(Q_s(t) - Q_k(t))x\|^2. \end{aligned}$$

Recalling that $Q_k^N(t) \geq Q_s^N(t)$, $Q_k(t) \geq Q_s(t)$ by construction, and using the uniform boundedness of Q_k^N and Q_s^N following from the uniform decay rate and the arguments above, we obtain for some $\hat{M}$

$$\begin{aligned} \|(Q_s^N(t) - Q_k^N(t))P_H^N x\|^2 &\leq 2\hat{M} < (Q_k^N(t) - Q_s^N(t))P_H^N x, P_H^N x >_H \\ \|(Q_s(t) - Q_k(t))x\|^2 &\leq 2\hat{M} < (Q_k(t) - Q_s(t))x, x >_H . \end{aligned}$$

Again using (2.7), we have

$$\begin{aligned} &< (Q_k^N(t) - Q_s^N(t))P_H^N x, P_H^N x >_H \\ &\quad \leq < (Q_k^N(t_k) - Q_s^N(t_k))S^N(t_k,t)P_H^N x, S^N(t_k,t)P_H^N x >_H, \end{aligned}$$

and

$$< (Q_k(t) - Q_s(t))x, x >_H \leq < (Q_k(t_k) - Q_s(t_k))S(t_k,t)x, S(t_k,t)x >_H .$$

Combining the above inequalities, we obtain

$$\begin{aligned} \|Q_s^N(t)P_H^N x - Q_s(t)x\|^2 \leq \|Q_k^N(t)P_H^N x - Q_k(t)x\|^2 \\ +4M\hat{M}(\|S^N(t_k,t)x\|^2 + \|S(t_k,t)x\|^2). \end{aligned}$$

Let k be large enough so that

$$\|S(t_k,t)x\| \leq \epsilon/(12M\hat{M}), \quad \|S^N(t_k,t)P_H^N x\| \leq \epsilon/(12M\hat{M}),$$

for all N. Then let N be large enough to obtain $\|Q_k^N(t)P_H^N x - Q_k(t)x\|^2 \leq \epsilon/3$. From the previous estimates we thus find $\|Q_s^N(t)P_H^N x - Q_s(t)x\|^2 \leq \epsilon$ which yields the desired results.

We note that if Hypothesis 3.4 holds and the uniform decay assumption in Theorem 3.1 is replaced by the following: there exists $\hat{t}$ such that for any $t \geq \hat{t}$ and for any $s \in [t_0, t_f]$, $\|S(t+s,s)x\| \leq \epsilon$, $\|S^N(t+s,s)P_H^N x\| \leq \epsilon$, for all N, then the convergence of Theorem 3.1 is uniform in the bounded interval $[t_0, t_f]$.

Theorem 3.1 is not very useful in practice, since the uniform decay assumption is difficult to verify directly. However, it does provide some insight and suggests more realistic conditions that might be verifiable. Recalling the definition of W-detectability and our discussions following it, we are prompted to formulate the following assumptions.

Hypothesis 3.6. (*Uniform W-detectability*). The original system is detectable and there exist constants M_K, M_{KV}, and $\beta > 0$ such that for each N, there exists a uniformly bounded operator valued function $K^N(\cdot) : H^N \mapsto H^N$, with $\|K^N(t)\|_{L(H^N)} \leq M_K$, for $t \in [t_0, \infty)$. If $T_{K^N}^N$ is the evolution operator corresponding to the perturbation of T^N by $K^N\sqrt{W^N}$, then $\|T_{K^N}^N(t,s)\|_{L(H^N)} \leq M_{KV}e^{-\beta(t-s)}$, for $(t,s) \in \triangle_\infty(t_0)$.

If Hypothesis 3.6 holds, then $Q_{\min}^N$ is the unique uniformly bounded solution of the Riccati integral equation on the infinite time interval for H^N. Under the uniform W-stabilizability Hypothesis 3.3, we have $Q_{\min}^N(t) \leq M \cdot I$, for all $t \in [t_0, \infty)$. Furthermore, by an application of Theorem 2.2, there exist constants $M_s, \alpha > 0$ independent of N such that the evolution operator S^N defined via $Q_{\min}^N$ satisfies

$$\|S^N(t,s)\| \leq M_s e^{-\alpha(t-s)}, \quad \text{for } (t,s) \in \triangle_\infty(t_0).$$

Thus, by Theorem 3.1, $Q^N_{\min}(t)$ converges to $Q_{\min}(t)$ as $N \to \infty$. We summarize the results in a major convergence theorem.

THEOREM 3.2. *Assume that system* (2.1), (2.2) *and its approximate systems* (3.1), (3.2) *satisfy Hypotheses* 3.1 *and* 3.2 *and the uniform* W*-stabilizability and uniform* W*-detectability conditions of Hypotheses* 3.3, 3.6. *Then the unique uniformly bounded solutions* Q^N *of the approximating Riccati integral equations* ((2.9) *with* T, B, W, R *replaced by* T^N, B^N, W^N, R^N) *on* $[t_0, \infty)$ *in* H^N *converge strongly to the unique uniformly bounded solution* Q *of the Riccati integral equation* (2.9) *on the infinite time interval in* H. *Furthermore, if Hypothesis* 3.2 *is replaced by Hypothesis* 3.4, *then this convergence is uniform in* t *for* t *in any bounded interval.*

We note that in the case of a periodic system, the uniform convergence in one period implies that Q^N converges to Q uniformly in the entire interval $[t_0, \infty)$. We further remark that the convergence of the Riccati operator guaranteed by Theorem 3.2 is sufficient (using standard arguments, see [G], [BK1]) to guarantee convergence of the optimal approximate feedback system trajectories $S^N(t,s)P^N_H x$ and optimal approximate controls u^N to the optimal system trajectories $S(t,s)x$ and optimal controls u (see Theorem 3.1 of [BK1]). Moreover, one also obtains convergence of the system generated by using the approximate feedback gains with the original infinite-dimensional control system (a feature that is of great practical importance), e.g., see the related remarks in §4 of [BK1].

The hypotheses of Theorem 3.2 are much more readily verified than others guaranteeing convergence that can be found in the literature (e.g., see [G, Thm. 5.3], where one is required to show that the approximate systems are uniformly stabilized by the feedback with a uniformly bounded sequence of approximate Riccati operators). As we shall see in the later sections, there are two distinct approaches that lead to rather easy use of our Theorem 3.2 in the event one is dealing with parabolic evolution systems.

4. Parabolic evolution equations: control and approximation. In this section, we formulate the linear quadratic regulator problem for an abstract parabolic control system. We focus our attention on systems associated with a time dependent sesquilinear form. First we review the theory of parabolic evolution equations (relying heavily on [T]) and extend some related results in a form applicable to control problems. Then a control system is defined for which general assumptions of stabilizability and detectability are made. A framework for approximation schemes is presented and conditions for convergence of the operators involved are discussed under assumptions of uniform stabilizability and uniform detectability for the approximate systems. Our discussions here are in the spirit of the approaches taken in [BK1], [BI1], [BI2].

Let H, V be two complex separable Hilbert spaces with $< \cdot, \cdot >_H$, $< \cdot, \cdot >_V$ as inner products and $\| \cdot \|_H$, $\| \cdot \|_V$ as norms, respectively. Let V^* be the dual space of V with $< \cdot, \cdot >_{V^*,V}$ denoting the duality pairing. The space V is assumed to be densely and continuously embedded in H, and thus there exists a constant c such that for all $\psi \in V$, $\|\psi\|_H \le c\|\psi\|_V$. Since for each element φ of H, we can define a bounded linear functional on V by $< \varphi, \psi >_H$, for $\psi \in V$, we have the usual embedding relationship $V \subset H \subset V^*$.

For each t in the interval $[t_0, \infty)$, consider a sesquilinear form $\sigma(t; \cdot, \cdot)$ defined on $V \times V$. We assume throughout that σ has the following properties:

Hypothesis 4.1. (V*-Continuity*). For each bounded interval $[t_0, t_1]$, there exists a

constant c_1 such that

$$|\sigma(t;\varphi,\psi)| \le c_1\|\varphi\|_V\|\psi\|_V, \quad \text{for } t \in [t_0,t_1], \varphi,\psi \in V. \tag{4.1}$$

Hypothesis 4.2. (*V-Ellipticity*). For each bounded interval $[t_0,t_1]$, there exist constants $c_2 > 0$, m such that

$$\text{Re } \sigma(t;\varphi,\varphi) \ge c_2\|\varphi\|_V^2 - m\|\varphi\|_H^2, \quad \text{for } t \in [t_0,t_1], \varphi \in V. \tag{4.2}$$

Under the above assumptions, we have a well-known ([FM], [K], [T], [S]) result: For each $t \in [t_0,t_1]$, there exists a unique closed operator $A(t) : V \mapsto V^*$ such that

$$\sigma(t;\varphi,\psi) = - < A(t)\varphi,\psi >_{V^*,V}, \quad \text{for } \psi \in V. \tag{4.3}$$

Furthermore, if $\hat{A}(t)$ is defined using the same method with a sesquilinear form σ^* defined by $\sigma^*(t;\varphi,\psi) = \overline{\sigma(t;\psi,\varphi)}$, then $\hat{A}(t)$ is identical to the adjoint operator $A^*(t)$ of $A(t)$. Both operators $A(t)$, $A(t)^*$ are infinitesimal generators of analytic semigroups in V^*, and an abstract parabolic evolution equation can be defined by

$$\frac{d}{dt}x(t) = A(t)x(t), \quad x(t_0) = x_0 \in V^*.$$

In order to insure the existence of an evolution operator for this equation, we must make additional assumptions on the continuity of σ with respect to t.

Hypothesis 4.3. (*Smoothness in t*). For each bounded interval $[t_0,t_1]$, there exist constants K and α, $0 < \alpha \le 1$, such that for all $t,s \in [t_0,t_1]$, and for all $\varphi,\psi \in V$, we have

$$|\sigma(t;\varphi,\psi) - \sigma(s;\varphi,\psi)| \le K|t-s|^\alpha\|\varphi\|_V\|\psi\|_V.$$

Under the above assumptions, there exists an evolution operator associated with the above evolution equation. The following theorem summarizes the properties of this evolution operator.

THEOREM 4.1. ([T, pp.127, pp.145-155]). *Let Hypotheses* 4.1–4.3 *hold. Then there exists a unique evolution operator* $\tilde{T}(\cdot,\cdot)$ *in* V^* *satisfying the following conditions:*

(i) *For any* $t_0 \le s < t \le t_1$, *the range* $\mathcal{R}(\tilde{T}(t,s))$ *of operator* $\tilde{T}(t,s)$ *is a subset of* V.

(ii) *The operator* $\tilde{T}(t,s)A(s)$ *has a unique bounded extension in* $L(V^*)$, *for all* $t_0 \le s < t \le t_1$; *therefore, we can and will use the same expression for the extension.*

(iii) *For each* $\varphi \in V^*$, *the* V^**-valued function* $\tilde{T}(t,s)\varphi$ *is continuously differentiable in* t *for* $t \in (s,t_1]$, *and continuously differentiable in* s *for* $s \in [t_0,t)$. *Furthermore, for* $\varphi \in V^*$

$$\begin{aligned} \frac{d}{dt}\tilde{T}(t,s)\varphi &= A(t)\tilde{T}(t,s)\varphi, \\ \frac{d}{ds}\tilde{T}(t,s)\varphi &= -\tilde{T}(t,s)A(s)\varphi. \end{aligned}$$

(iv) *The restriction of* $\tilde{T}(t,s)$ *on* H *is strongly continuous in the* H *norm. For all* $x_0 \in H$, *the function* $x(t) = \tilde{T}(t,s)x_0$ *is in* $L^2([s,t_1];V)$ *and the derivative* $\dot{x}(t) = A(t)\tilde{T}(t,s)x_0$ *is in* $L^2([s,t_1];V^*)$. *Furthermore, there exist constants* C_1, C_2, *depending only on* c_1, c_2, m, K, *and* α *such that*

$$\|\tilde{T}(t,s)x_0\|_V \leq C_1(t-s)^{-1/2}\|x_0\|_H. \tag{4.4}$$

$$\|\tilde{T}(\cdot,s)x_0\|_{L^2([s,t_1];V)} \leq C_2\|x_0\|_H. \tag{4.5}$$

All the statements in the above theorem can be found in [T]. However, they are organized into several sections with somewhat different notation; we therefore give a brief argument which collects the results from the book.

Proof. Existence. Taking $X = V^*$, we let $A(t)$ be defined as in (4.3). As indicated in [T, p. 144], using Theorem 5.2.1 of [T], we find there exists an evolution operator $\tilde{T}$ on V^*. The range $\mathcal{R}(\tilde{T}(t,s))$ is a subset of $\mathcal{D}(A(t)) = V$ for all $t_0 \leq s < t \leq t_1$. For any $\varphi \in V^*$, $\tilde{T}(t,s)\varphi$ is continuously differentiable in t for $t \in (s,t_1]$. Now let the sesquilinear form σ^* be defined by

$$\sigma^*(t;\varphi,\psi) = \overline{\sigma(t;\psi,\varphi)}, \quad \varphi,\psi \in V.$$

Let $A^*(t)$ be the linear operator defined via σ^*; then $A^*(t)$ is the adjoint operator of $A(t)$. As indicated by the remarks following Lemma 5.4.6 of [T], we can use the results of §5.2 (with $\tilde{S}$ and A^* replacing U and A of [T]) to construct an operator-valued function $\tilde{S}(t,s)$, such that for all $t_0 \leq s < t \leq t_1$, $A^*(s)\tilde{S}(t,s)$ is a bounded operator in V^*, and for any $\varphi \in V^*$, $\tilde{S}(t,s)\varphi$ is continuously differentiable in s for $s \in [t_0,t)$. Furthermore, for $\varphi \in V^*$

$$\frac{d}{ds}\tilde{S}(t,s)\varphi = -A^*(s)\tilde{S}(t,s)\varphi.$$

In fact, $\tilde{S}(t,s)$ can be constructed as follows. Let $\exp\{tA^*(s)\}$ be the semigroup generated by $A^*(s)$ and we define

$$\begin{aligned} \tilde{S}(t,s) &= \exp\{(t-s)A^*(t)\} + W(t,s), \\ W(t,s) &= \int_s^t \exp\{(\tau-s)A^*(\tau)\}R(t,\tau)d\tau, \end{aligned}$$

where the function R can be computed by iterative methods using

$$R(t,s) - \int_s^t R_1(\eta,s)R(t,\eta)d\eta = R_1(t,s),$$

with $R_1(t,s) = (A^*(t) - A^*(s))\exp\{(t-s)A^*(t)\}$. Then following the same type of arguments as in [T, p. 149], we can conclude that $\tilde{S}(t,s) = \tilde{T}^*(t,s)$. Therefore, $\tilde{T}(t,s)A(s)$ has a unique bounded extension in V^*. For all $\varphi \in V^*$, $\tilde{T}(t,s)\varphi$ is strongly differentiable in s for $s \in [t_0,t)$.

Finally, statement (iv) of the above theorem can be found in the §§5.4 and 5.5 of [T]. We note that in these sections of [T], the space X plays the role of our space H. Let $T(t,s)$ be the restriction of $\tilde{T}(t,s)$ to H; by Theorem 5.4.1 of [T], $T(t,s)$ is strongly continuous in the H norm. Furthermore the estimate (4.4) holds. For

any $x_0 \in H$, let $x(t) = T(t,s)x_0$. By Lemma 5.5.2 and Proposition 5.5.1 of [T, pp. 152–153] with $f \equiv 0$, the function $x(\cdot)$ is in $L^2([s,t_1];V)$ and $\dot{x}(\cdot)$ is in $L^2([s,t_1];V^*)$. In addition, the estimate (4.5) holds. We note that the constants C_1, C_2 depend only on the constants c_1, c_2, m, K, and α.

Uniqueness. By Theorem 5.2.3 of [T, p.128], we can conclude that the evolution operator satisfying the conditions (i)–(iv) must be unique.

We remark that the same theorem holds if we use the sesquilinear form σ^*; therefore, properties (i)–(iv) hold for the adjoint evolution operator $\tilde{T}^*(\cdot,\cdot)$. As a consequence of (iv), the restriction T of $\tilde{T}$ to H is an evolution operator in H as defined in § 2. We wish to take H as our state space since it is in this Hilbert space that our control problems are defined and our subsequent computational considerations are readily pursued; therefore, we use primarily the evolution operator T in this paper. The operator $\tilde{T}$ is used in the remainder of the current section in several proofs of uniqueness theorems. The only precaution one must take is that $T(t,s)\varphi$ is continuously differentiable with respect to t in the V^* sense and the derivative of $T(t,s)\varphi$ is an element of V^*. In particular, for each $\psi \in V$, $< T(t,s)\varphi, \psi >_H$ is differentiable with respect to t, and

$$\frac{d}{dt} < T(t,s)\varphi, \psi >_H = < A(t)T(t,s)\varphi, \psi >_{V^*,V} = -\sigma(t; T(t,s)\varphi, \psi).$$

The conclusions of this theorem are very useful in defining our control system. However, the conditions of Hypothesis 4.3 are too restrictive for our use, since we may need to perturb the equation with nonsmooth but bounded (feedback) terms. We can show that if σ is perturbed with a sesquilinear form that is uniformly bounded in H, then there exists an associated evolution operator T_K which preserves most of the desirable properties of the evolution operator T. In fact, let $K(\cdot) : [t_0, \infty) \mapsto L(H)$ be a uniformly bounded operator valued measurable function. We can then define a sesquilinear form σ_K in $V \times V$ as

$$\sigma_K(t; \varphi, \psi) = \sigma(t; \varphi, \psi) - < K(t)\varphi, \psi >_H, \quad \varphi, \psi \in V.$$

It is easy to see that for each bounded interval $[t_0, t_1]$, Hypotheses 4.1 and 4.2 hold. Therefore, we can find an operator $A_K(t)$ defined on V such that (4.3) holds for σ_K and $A_K(t)$. Furthermore, we have by the definitions of $A(t)$ and $A_K(t)$ that $A_K(t)\varphi = A(t)\varphi + K(t)\varphi$ for $\varphi \in V$ and we may establish the following result.

THEOREM 4.2. *Consider a sesquilinear form σ satisfying Hypotheses* 4.1–4.3 *and let $K(\cdot)$, σ_K be defined as above. Then there exists a unique evolution operator $T_K(\cdot,\cdot)$ in H for which the following properties hold:*

(i) *The range $\mathcal{R}(T_K(t,s))$ of the operator $T_K(t,s)$ is a subset of V, for all $t_0 \leq s < t \leq t_1$.*

(ii) *For $\varphi \in H$, the function $T_K(t,s)\varphi$ is differentiable with respect to t in the V^* sense, and*

$$\frac{d}{dt} T_K(t,s)\varphi = A_K(t)T_K(t,s)\varphi.$$

(iii) *For all $x_0 \in H$, the function $x(t) = T_K(t,s)x_0$ is in $L^2([s,t_1];V)$ and its derivative $\dot{x}(t)$ is in $L^2([s,t_1];V^*)$. Furthermore, there exists constants C_1, C_2 depending only on c_1, c_2, m, K, and α such that*

$$\|T_K(t,s)x_0\|_V \leq C_1(t-s)^{-1/2}\|x_0\|_H,$$
$$\|T_K(\cdot,s)x_0\|_{L^2([s,t_1];V)} \leq C_2\|x_0\|_H.$$

Proof. Existence. Let T_K be the unique evolution operator in H corresponding to the perturbation of T by K. From the results on perturbations given in § 2, we have that T_K satisfies for all $\varphi \in H$

$$T_K(t,s)\varphi = T(t,s)\varphi + \int_s^t T(t,\eta)K(\eta)T_K(\eta,s)\varphi d\eta, \tag{4.6}$$

$$T_K(t,s)\varphi = T(t,s)\varphi + \int_s^t T_K(t,\eta)K(\eta)T(\eta,s)\varphi d\eta.$$

Since the function $K(\eta)T_K(\eta,s)\varphi$ is uniformly bounded in H norm by some constant C, using the estimate (4.4), we can find a constant $\tilde{C}$ such that for $\eta \in [s,t]$

$$\|T(t,\eta)K(\eta)T_K(\eta,s)\varphi\|_V \leq \tilde{C}(t-\eta)^{-1/2}\|\varphi\|_H.$$

Therefore, the integral term in (4.6) converges in the V sense and hence, $T_K(t,s)\varphi \in V$ for $\varphi \in H$.

Letting $x_0 \in H$, we define $x(t) = T_K(t,t_0)x_0$, and $f(t) = K(t)x(t)$. From (4.6), the function $x(t)$ can be written as

$$x(t) = T(t,t_0)x_0 + \int_{t_0}^t T(t,\eta)f(\eta)d\eta.$$

By the strong continuity of T_K and uniform boundedness of K, it is obvious that $f(\cdot) \in L^2([t_0,t_1];H)$ and hence $f(\cdot) \in L^2([t_0,t_1];V^*)$. By Theorem 5.5.1 of [T], $x(\cdot)$ is in $L^2([t_0,t_1];V)$, is differentiable with $\dot{x}(\cdot)$ in $L^2([t_0,t_1];V^*)$ and satisfies $\dot{x}(t) = A_K(t)x(t)$. Using the equality (4.6) and the boundedness of the perturbation, by modifying the constants C_1, C_2 in (4.4), (4.5), we can easily obtain

$$\|T_K(t,s)\varphi\|_V \leq C_1(t-s)^{-1/2}\|\varphi\|_H,$$
$$\|T_K(\cdot,t_0)\varphi\|_{L^2([t_0,t_1];V)} \leq C_2\|\varphi\|_H,$$

for all $\varphi \in H$.

Uniqueness. Let $\hat{T}_K$ satisfy the conclusions (i)–(ii) of Theorem 4.2. For all $\varphi \in H$, consider $\hat{T}_K(t,s)\varphi$ as a V^* valued function. Then we have

$$\begin{aligned}\frac{d}{d\eta}\tilde{T}(t,\eta)\hat{T}_K(\eta,s)\varphi &= -\tilde{T}(t,\eta)A(\eta)\hat{T}_K(\eta,s)\varphi \\ &\quad +\tilde{T}(t,\eta)(A(\eta)+K(\eta))\hat{T}_K(\eta,s)\varphi \\ &= \tilde{T}(t,\eta)K(\eta)\hat{T}_K(\eta,s)\varphi.\end{aligned}$$

Integrating both sides of the above equation from s to t, we obtain

$$\hat{T}_K(t,s)\varphi = \tilde{T}(t,s)\varphi + \int_s^t \tilde{T}(t,\eta)K(\eta)\hat{T}_K(\eta,s)\varphi d\eta.$$

Since T is the restriction of $\tilde{T}$ to H, $\hat{T}_K$ is a solution of (4.6). By the uniqueness results of § 2 for perturbed evolution operators, we have $\hat{T}_K = T_K$. Hence, the unique solution of (4.6) is the unique evolution operator T_K generated in the theorem.

Now consider a function $f(\cdot) \in L^2([t_0, t_1]; H)$. We can then define

$$z(t) = T(t, t_0)z_0 + \int_{t_0}^{t} T(t, \eta)f(\eta)d\eta.$$

The function $z(\cdot)$ is the unique solution of the following initial value problem:

$$\begin{aligned} z(t) &= T(t, s)z(s) + \int_s^t T(t, \eta)f(\eta)d\eta, \quad t_0 \le s \le t \le t_1. \\ z(t_0) &= z_0. \end{aligned} \tag{4.7}$$

Henceforth, we consider (4.7) as the definition of our basic evolution system. The function z corresponds to the solution of a weaker formulation of the evolution equation.

LEMMA 4.1. ([T, Thm. 5.5.1]). *The function $z(\cdot)$ given by* (4.7) *is the unique function in $L^2([t_0, t_1]; V)$ with derivative $\dot{z}(\cdot)$ in $L^2([t_0, t_1]; V^*)$ for which the following equation holds for $\psi \in V$*

$$\begin{aligned} < z(t) - z(t_0), \psi >_H &= \int_{t_0}^{t} \{-\sigma(\eta; z(\eta), \psi) + < f(\eta), \psi >_H\} d\eta, \\ z(t_0) &= z_0. \end{aligned} \tag{4.8}$$

LEMMA 4.2. ([T, Lemma 5.5.1]). *For any two functions $z(\cdot), w(\cdot)$ in $L^2([t_0, t_1]; V)$ with derivatives $\dot{z}, \dot{w}$ in $L^2([t_0, t_1]; V^*)$, the following equality holds:*

$$\begin{aligned} < z(t), w(t) >_H &= < z(s), w(s) >_H \\ &\quad + \int_s^t \left\{ < \dot{z}(\eta), w(\eta) >_{V^*, V} + \overline{< \dot{w}(\eta), z(\eta) >_{V^*, V}} \right\} d\eta, \end{aligned}$$

for all $t_0 \le s \le t \le t_1$.

As a consequence of these lemmas, if for any $x_0 \in H$, we let $x(t) = T(t, t_0)x_0$, then

$$\|x(t)\|_H^2 = \|x_0\|_H^2 - 2\int_{t_0}^{t} \operatorname{Re}\, \sigma(\eta; x(\eta), x(\eta))d\eta. \tag{4.9}$$

We note that if Hypotheses 4.1–4.3 hold, then for each bounded interval $[t_0, t_1]$, we can define $T(t, s)$ uniquely. Therefore, $T(t, s)$ is also uniquely defined for all $t_0 \le s \le t < \infty$. The equality (4.9) suggests a sufficient condition for the stability of T.

Hypothesis 4.4. There exists a constant $k > 0$ such that

$$\operatorname{Re}\, \sigma(t; \varphi, \varphi) \ge k\|\varphi\|_H^2, \quad \text{for } t_0 \le t < \infty, \quad \varphi \in V.$$

THEOREM 4.3. *In addition to Hypotheses* 4.1–4.3, *under Hypothesis* 4.4, *T is uniformly exponentially stable.*

Proof. For any $x_0 \in H$, let $x(t) = T(t, s)x_0$. Then by (4.9), we have

$$\|x(t)\|_H^2 \le \|x_0\|_H^2 - 2\int_s^t k\|x(\eta)\|^2 d\eta,$$

for all $t_0 \leq s \leq t < \infty$. This implies

$$\|T(t,s)\|_{L(H)} \leq 1, \quad \int_s^t \|T(\eta,s)x_0\|_H^2 \leq \frac{1}{2k}\|x_0\|_H^2.$$

Therefore, by Lemma 2.1, T is uniformly exponentially stable. Note, moreover, that by Lemma 2.1, under Hypothesis 4.4, we can find $M, \alpha > 0$ depending only on k such that

$$\|T(t,s)\|_{L(H)} \leq Me^{-\alpha(t-s)}.$$

We can now use these considerations to define an evolution equation control system of the form (4.7) via a sesquilinear form. The space H will serve as our state space, with subspace V and the sesquilinear form σ defined as above and Hypotheses 4.1–4.3 holding. Let the control space U be a Hilbert space, and let $B(\cdot) : [t_0, \infty) \mapsto L(U,H)$ be a strongly measurable operator-valued function. We assume that there exists a constant M_B such that

$$\|B(t)\|_{L(U,H)} \leq M_B, \quad \text{for } t \in [t_0, \infty).$$

For any control $u(\cdot) : [t_0, \infty) \mapsto U$, belonging to $L^2([t_0,\infty);U)$, the corresponding trajectories satisfy for $\psi \in V$

$$(4.10) \qquad < z(t) - z(s), \psi >_H = -\int_s^t \{\sigma(\eta; z(\eta), \psi) - < B(\eta)u(\eta), \psi >_H\}\, d\eta,$$

for all $(t,s) \in \triangle_\infty(t_0)$. Let $T(\cdot,\cdot)$ be the evolution operator defined via σ. By Lemma 4.1, an equivalent form of (4.10) is given by

$$(4.11) \qquad z(t) = T(t,s)z(s) + \int_s^t T(t,\eta)B(\eta)u(\eta)d\eta, \quad \text{for } (t,s) \in \triangle_\infty(t_0).$$

Let $z_0 \in H$ be the initial state of the system at t_0 and let the cost for control $u(\cdot)$ be given by

$$(4.12) \qquad J_\infty(u; z_0, t_0) = \int_{t_0}^\infty < W(t)z(t), z(t) >_H + < R(t)u(t), u(t) >_U \, dt,$$

where $W(\cdot) : [t_0,\infty) \mapsto L(H)$, $R(\cdot) : [t_0,\infty) \mapsto L(U)$ are strongly measurable. The operators $W(t), R(t)$ are assumed to be self-adjoint nonnegative definite operators, uniformly bounded in the entire interval $[t_0,\infty)$. Furthermore, there exists a constant $r > 0$ such that

$$< R(t)v, v >_U \quad \geq \quad r\|v\|_U^2, \quad \text{for } t \geq t_0, \quad v \in U.$$

Recalling the discussions of § 2, we note that the standing assumptions of that section hold. Therefore, for a given nonnegative definite self-adjoint operator G on H, the Riccati integral equation in each finite time interval $[t_0, t_k]$,

$$\begin{aligned} Q_k(s)x \quad &= \quad T^*(t_k,s)GT(t_k,s)x \\ &\quad + \int_s^{t_k} T^*(\eta,s)\left[W(\eta) - Q_k(\eta)B(\eta)R^{-1}(\eta)B^*(\eta)Q_k(\eta)\right]T(\eta,s)x d\eta \end{aligned}$$

has a unique self-adjoint solution Q_k.

For the control problem in the infinite time interval $[t_0, \infty)$, we need stabilizability and detectability conditions to assure existence and uniqueness of a uniformly bounded solution of the Riccati integral equation.

Hypothesis 4.5. (*Detectability*). There exists a strongly measurable uniformly bounded operator valued function $\Psi(\cdot) : [t_0, \infty) \mapsto L(H)$, such that if we denote by S_Ψ the evolution operator corresponding to the perturbation of T by $\Psi(\cdot)W^{1/2}(\cdot)$, the following estimate holds for $x \in H$:

$$\|S_\Psi(t,s)x\|_H \leq Me^{-\omega(t-s)}\|x\|_H,$$

for some constants $M, \omega > 0$.

Hypothesis 4.6. (*Stabilizability*). There exists a strongly measurable uniformly bounded operator valued function $K(\cdot) : [t_0, \infty) \mapsto L(H, U)$, such that if we denote by S_K the evolution operator corresponding to the perturbation of T by $B(\cdot)K(\cdot)$, the following estimate holds for $x \in H$:

$$\|S_K(t,s)x\|_H \leq Me^{-\omega(t-s)}\|x\|_H,$$

for some constants $M, \omega > 0$.

We remark that Hypothesis 4.6 is stronger than "W-stabilizability"; however, under Hypothesis 4.5, by Theorem 2.2, these two types of stabilizability assumption are equivalent.

To this point we have defined a control system using an abstract parabolic evolution equation that fits into the general framework of § 3. Under Hypotheses 4.5 and 4.6, we may apply the theory of the previous sections to establish the following results for our control problem:

(i) The Riccati integral equation in the infinite time interval $[t_0, \infty)$ has a unique uniformly bounded solution $Q(\cdot)$.

(ii) Let S_Q be the evolution operator corresponding to the perturbation of $T(\cdot,\cdot)$ by $-BR^{-1}B^*Q(\cdot)$. For each initial state z_0, the unique optimal trajectory is given by $S_Q(t, t_0)z_0$.

We turn next to giving results for finite-dimensional approximations of our control system. As in § 3, let $\{H^N\}_{N=1}^\infty$ be a sequence of finite-dimensional subspaces of $V \subset H$. Let P_H^N be the orthogonal projection operator from H onto H^N. Since H^N is an approximation of H, we assume that for every $\varphi \in H$, $\|P_H^N\varphi - \varphi\|_H \to 0$, as $N \to \infty$. In addition, we require that H^N is an approximation of V as well, so that for all $\varphi \in V$, $\|P_H^N\varphi - \varphi\|_V \to 0$, as $N \to \infty$. We note that in fact this latter convergence implies the convergence in H for $\varphi \in H$ since V is continuously and densely embedded in H.

Let $\{U^N\}_{N=1}^\infty$ be a sequence of finite-dimensional subspaces of U. Let P_U^N be the orthogonal projection operator from U onto U^N. We assume U^N approximates U in the following sense: for $v \in U$, $\|P_U^N v - v\|_U \to 0$, as $N \to \infty$.

For each N, we define a sesquilinear form σ^N as the restriction of σ to $H^N \times H^N$ and define a linear operator $A^N(t) : H^N \mapsto H^N$ by

$$- < A^N(t)\varphi^N, \psi^N >_H = \sigma^N(t; \varphi^N, \psi^N), \quad \text{for } \varphi^N, \psi^N \in H^N.$$

By continuity of σ with respect to t, the operator valued function $A^N(t)$ is continuous in time. As a consequence, there exists a unique differentiable evolution operator $T^N(\cdot,\cdot)$ in H^N generated by $A^N(t)$; that is, for $\varphi^N \in H^N$ we have

$$\frac{d}{dt}T^N(t,s)\varphi^N = A^N(t)T^N(t,s)\varphi^N.$$

Note immediately that the σ^N's satisfy Hypotheses 4.1–4.3 with the same constants c_1, c_2, m, α, and K; therefore, for each fixed interval $[t_0,t_1]$, there exist constants C_1, C_2 independent of N such that for all $\varphi^N \in H^N$,

$$\|T^N(t,s)\varphi^N\|_V \le C_1(t-s)^{-\frac{1}{2}}\|\varphi^N\|_H; \tag{4.13}$$

$$\|T^N(\cdot,t_0)\varphi^N\|_{L^2([t_0,t_1];V)} \le C_2\|\varphi^N\|_H. \tag{4.14}$$

The approximation properties of the evolution operator T^N are summarized by the the following convergence theorem.

THEOREM 4.4. *Let Hypotheses* 4.1–4.3 *hold and let* $T(\cdot,\cdot)$ *and* $T^N(\cdot,\cdot)$ *be defined as above where* $\|P_H^N\varphi - \varphi\|_V \to 0$ *as* $N \to \infty$ *for* $\varphi \in V$*; then the following properties hold:*

(i) *There exist constants* M_T *and* ω *such that for all* N,

$$\|T^N(t,s)\|_{L(H^N)} \le M_T e^{\omega(t-s)}, \|T(t,s)\|_{L(H)} \le M_T e^{\omega(t-s)}, \quad t_0 \le s \le t < \infty.$$

(ii) *For any finite interval* $[a,b] \subset [t_0,\infty)$ *and any* $\varphi \in H$, *we have*

$$\|T^N(t,s)P_H^N\varphi - T(t,s)\varphi\|_H \to 0, \quad \text{as } N \to \infty, \quad a \le s \le t \le b.$$

Furthermore, the convergence is uniform for all $a \le s \le t \le b$.

Proof. (i) By Lemma 4.2 and (4.9), for every $\varphi \in H^N$ we have

$$\|T^N(t,s)\varphi\|_H^2 = \|\varphi\|_H^2 - 2\int_s^t \operatorname{Re}\, \sigma^N(\eta; T^N(\eta,s)\varphi, T^N(\eta,s)\varphi)d\eta.$$

Under Hypothesis 4.2,

$$\|T^N(t,s)\varphi\|_H^2 \le \|\varphi\|_H^2 + 2\int_s^t m\|T^N(\eta,s)\varphi\|_H^2 d\eta.$$

Using Gronwall's inequality, we obtain

$$\|T^N(t,s)\varphi\|_H^2 \le \|\varphi\|_H^2 e^{2|m|(t-s)}.$$

Noting that the same estimates hold for $T(t,s)\varphi$, we obtain (i).

(ii) Let $\varphi \in H$, define $w(t) = T(t,s)\varphi$ and $w^N(t) = T^N(t,s)P_H^N\varphi$, and let $z^N(t) = w(t) - w^N(t)$. We note that $z^N(t)$ is not an element of H^N; in fact

$$z^N(t) - P_H^N z^N(t) = w(t) - P_H^N w(t). \tag{4.15}$$

Since $w(t)$ is differentiable in the V^* sense, $w^N(t)$ is differentiable, and both functions are in $L^2([a,b];V)$ with derivatives in $L^2([a,b];V^*)$. By (4.9), Lemma 4.2, and

definitions of the operators $A(t), A^N(t)$, we obtain

$$\begin{aligned}
&\|z^N(t)\|_H^2 \\
&\quad = \|z^N(0)\|_H^2 + 2\int_s^t \text{Re } < A(\eta)w(\eta) - A^N(\eta)w^N(\eta), z^N(\eta) >_{V^*,V} d\eta \\
&\quad = \|z^N(0)\|_H^2 - 2\int_s^t \text{Re } \{\sigma(\eta; w(\eta), z^N(\eta)) - \sigma^N(\eta; w^N(\eta), P_H^N z^N(\eta))\} d\eta \\
&\quad -2\int_s^t < A^N(\eta)w^N(\eta), z^N(\eta) - P_H^N z^N(\eta) >_{V^*,V} d\eta.
\end{aligned}$$

Since the duality pairing reduces to the H-inner product on $H \times H$, we have

$$\begin{aligned}
&< A^N(\eta)w^N(\eta), z^N(\eta) - P_H^N z^N(\eta) >_{V^*,V} \\
&\quad =< A^N(\eta)w^N(\eta), z^N(\eta) - P_H^N z^N(\eta) >_H .
\end{aligned}$$

Moreover, P_H^N is the orthogonal projection operator and hence the last term in the above equation equals to zero. Using the definition of σ^N, the sesquilinearity of σ and (4.15), we find

$$\begin{aligned}
\|z^N(t)\|_H^2 &= \|\varphi - P_H^N\varphi\|_H^2 - 2\int_s^t \text{Re } \sigma(\eta; w(\eta), z^N(\eta) - P_H^N z^N(\eta))d\eta \\
&\quad -2\int_s^t \text{Re } \sigma(\eta; w(\eta), P_H^N z^N(\eta))d\eta \\
&\quad +2\int_s^t \text{Re } \sigma^N(\eta; w^N(\eta), P_H^N z^N(\eta))d\eta \\
&= \|\varphi - P_H^N\varphi\|_H^2 - 2\int_s^t \text{Re } \sigma(\eta; w(\eta), w(\eta) - P_H^N w(\eta))d\eta \\
&\quad -2\int_s^t \text{Re } \sigma(\eta; w(\eta) - P_H^N w(\eta), P_H^N z^N(\eta))d\eta \\
&\quad -2\int_s^t \text{Re } \sigma^N(\eta; P_H^N z^N(\eta), P_H^N z^N(\eta))d\eta.
\end{aligned}$$

Since P_H^N is the orthogonal projection, we find $< P_H^N z^N, w - P_H^N w >_H = 0$, so that from (4.15)

$$\|P_H^N z^N(t)\|_H^2 = \|z^N(t)\|_H^2 - \|w(t) - P_H^N w(t)\|_H^2.$$

Combining this with the previous equation, we have

$$\|P_H^N z^N(t)\|_H^2 = \Theta^N(t,s) - 2\int_s^t \text{Re } \sigma^N(\eta; P_H^N z^N(\eta), P_H^N z^N(\eta))d\eta,$$

where $\Theta^N(t,s)$ is given by

$$\begin{aligned}
\Theta^N(t,s) &= \|\varphi - P_H^N\varphi\|_H^2 - \|w(t) - P_H^N w(t)\|_H^2 \\
&\quad -2\int_s^t \text{Re } \sigma(\eta; w(\eta), w(\eta) - P_H^N w(\eta))d\eta \\
&\quad -2\int_s^t \text{Re } \sigma(\eta; w(\eta) - P_H^N w(\eta), P_H^N z^N(\eta))d\eta.
\end{aligned}$$

Using the V-ellipticity of σ^N, we find

$$\|P_H^N z^N(t)\|_H^2 \leq |\Theta^N(t,s)| + 2|m| \int_s^t \|P_H^N z^N(\eta)\|_H^2 d\eta. \tag{4.16}$$

To use Gronwall's inequality to conclude convergence of $P_H^N z^N$, it suffices to show that $|\Theta^N(t,s)|$ goes to zero uniformly for all $a \leq s \leq t \leq b$. By the continuity and uniform boundedness of T, the term $\|T(t,s)\varphi - P_H^N T(t,s)\varphi\|_H^2$ goes to zero uniformly for all $a \leq s \leq t \leq b$. Using the V-continuity of σ, the two integrals in Θ can be bounded by

$$2c_1 \int_s^t \left\{ \|P_H^N w(\eta) - w(\eta)\|_V \|P_H^N \dot{z}^N(\eta)\|_V + \|w(\eta)\|_V \|P_H^N w(\eta) - w(\eta)\|_V \right\} d\eta$$

$$\leq 2c_1 \left[\|w(\cdot)\|_{L^2([a,b];V)} + \|P_H^N z^N(\cdot)\|_{L^2([a,b];V)} \right] \left[\int_s^t \|w(\eta) - P_H^N w(\eta)\|_V^2 d\eta \right]^{\frac{1}{2}}.$$

Using the inequalities (4.5) and (4.14), we observe that the functions w, w^N are in a bounded subset of $L^2([a,b];V)$. By dominated convergence arguments, the above integral converges to zero. Furthermore, by taking $t = b$, $s = a$, we obtain that this convergence is uniform for all $a \leq s \leq t \leq b$. Therefore, $|\Theta^N(t,s)|$ converges to zero uniformly for all $a \leq s \leq t \leq b$. Finally, from (4.15), we have

$$\|T(t,s)\varphi - T^N(t,s)P_H^N\varphi\|_H^2 = \|P_H^N z^N(t)\|_H^2 + \|w(t) - P_H^N w(t)\|_H^2,$$

and the uniform convergence of $P_H^N z^N(t)$ in t implies $T^N(t,s)P^N\varphi$ converges to $T(t,s)\varphi$ uniformly for all $a \leq s \leq t \leq b$.

Since we can define operators $A^*(t), A^{*N}, T^*(t,s)$ and $T^*(t,s)$ by using the sesquilinear form σ^* as we indicated after (4.3), the convergence of $T^{*N}(t,s)$ to $T^*(t,s)$ can be shown using the same arguments as in the proof of the above theorem.

Having defined our approximate (uncontrolled) system and established the convergence of Theorem 4.4, we return to the control problem for (4.10)–(4.12). Approximations of functions B, W, R are defined as follows:

$$\begin{array}{lll} B^N(\cdot) : [t_0,\infty) \mapsto L(U^N, H^N), & B^N(t)v^N = P_H^N B(t)v^N, & v^N \in U^N; \\ W^N(\cdot) : [t_0,\infty) \mapsto L(H^N), & W^N(t)\varphi^N = P_H^N W(t)\varphi^N, & \varphi^N \in H^N; \\ R^N(\cdot) : [t_0,\infty) \mapsto L(U^N), & R^N(t)v^N = P_U^N R(t)v^N, & v^N \in U^N. \end{array}$$

Let G be the nonnegative self-adjoint operator in the finite interval cost functional associated in the usual manner with (4.12) for our control system in H. Let $G^N = P_H^N G$ and $z_0^N = P_H^N z_0$.

In each subspace H^N, a finite-dimensional control system is thus defined by

$$z^N(t) = T^N(t,s)z^N(s) + \int_s^t T^N(t,\eta)B^N(\eta)u^N(\eta)d\eta, \tag{4.17}$$

with $u^N(\cdot) \in L^2([t_0,\infty);U^N)$ and $z^N(t_0) = z_0^N$. The cost functionals for the associated finite time interval problems are given by

$$J^N(u^N; z_0^N, t_0, t_k) = < G^N z^N(t_k), z^N(t_k) >_H + \int_{t_0}^{t_k} \left\{ < W^N(t)z^N(t), z^N(t) >_H + < R^N(t)u^N(t), u^N(t) >_U \right\} dt, \tag{4.18}$$

while the cost functional for the infinite time interval problem is given by

$$J_\infty^N(u^N; z_0^N; t_0) = \int_{t_0}^{\infty} \{< W^N(t)z^N(t), z^N(t) >_H + < R^N(t)u^N(t), u^N(t) >_U\} dt. \tag{4.19}$$

To obtain the uniform convergence of the operator valued functions, we make additional assumptions on the continuity of B, W, R.

Hypothesis 4.7. (*Parameter smoothness*). The operator valued functions B, W, R are piecewise strongly continuous functions on $[t_0, \infty)$.

LEMMA 4.3. *Under Hypothesis* 4.7, *the following convergence is uniform in* t *for* t *in any bounded interval:*

$$\begin{aligned}
&\|B^N(t)P_U^N v - B(t)v\|_H \to 0, && v \in U;\\
&\|B^{*N}(t)P_H^N \varphi - B^*(t)\varphi\|_U \to 0, && \varphi \in H;\\
&\|W^N(t)P_H^N \varphi - W(t)\varphi\|_H \to 0, && \varphi \in H;\\
&\|R^N(t)P_U^N v - R(t)v\|_U \to 0, && v \in U,
\end{aligned}$$

as $N \to \infty$. *The operator* $G^N P_H^N$ *also converges strongly to* G *as* $N \to \infty$.

Proof. We only prove the uniform convergence of B^N; the remainder of the arguments are similar. For simplicity, we without loss of generality assume that the function $t \mapsto B(t)$ is strongly continuous. For a given $v \in U$, the pointwise convergence of the functions $B^N(t)P_U^N v$ to $B(t)v$ is given by our assumptions on the approximation properties of the spaces H^N, U^N. To conclude uniform convergence in t, it is enough to show that the functions $B^N(t)P_U^N v$ are equicontinuous. That is, for any $\epsilon > 0$, there exists $\delta > 0$ such that for all N, if $|t - s| \leq \delta$, we have $\|B^N(t)P_U^N v - B^N(s)P_U^N v\|_H \leq \epsilon$. By definition of B^N, we have

$$\begin{aligned}
\|B^N(t)P_U^N v - B^N(s)P_U^N v\|_H &\leq \|B(t)P_U^N v - B(s)P_U^N v\|_H\\
&\leq \|B(t)(P_U^N v - v)\|_H + \|B(t)v - B(s)v\|_H\\
&\quad + \|B(s)(P_U^N v - v)\|_H\\
&\leq 2M_B\|P_U^N v - v\|_U + \|B(t)v - B(s)v\|_H.
\end{aligned}$$

By continuity of B, we conclude that $B^N(t)P_U^N v$ are equicontinuous functions of t in a bounded interval. Hence the convergence is uniform in any bounded interval.

It is easy to verify that B^N, W^N, R^N are uniformly bounded and W^N, R^N are nonnegative self-adjoint operators. In addition, there exists a constant $r > 0$ such that for all N,

$$< R^N(t)v^N, v^N >_U \geq r\|v^N\|_U^2, \quad \text{for } t \in [t_0, \infty), v^N \in U^N.$$

Consider any finite time interval $[t_0, t_k]$, and let Q_k, Q_k^N be the unique self-adjoint solutions of the Riccati integral equations in H and H^N associated with the control systems (4.11), (4.17), respectively. Then it follows from Theorem 4.4 and the discussions of §3 (in particular, see Hypotheses 3.1, 3.2, and the remarks just prior to Hypothesis 3.4) that for each k, $Q_k^N(t)$ converges to $Q_k(t)$ strongly and the convergence is uniform in t for $t \in [t_0, t_k]$.

We assumed above (Hypotheses 4.5, 4.6) that the control system (4.10)–(4.12) in H is detectable and stabilizable. Therefore, there exists a unique uniformly bounded solution Q of the Riccati integral equation on the infinite time interval $[t_0, \infty)$. In order to approximate Q by a uniformly bounded solution of the Riccati integral equation in H^N, we show that the approximate control systems defined here are also detectable and stabilizable. More importantly, recalling the results (e.g., see Theorem 3.2) of §3, to apply our results, we need uniform detectability and uniform stabilizability for the approximate systems (4.17), (4.19).

Based on the stabilizability and detectability properties of the original system, for a given approximation scheme, we would like to show that the following conditions hold.

Condition US (*Uniform stabilizability*). There exist constants $M_K, M, \omega > 0$ independent of N such that for each of the approximate systems, we can find a uniformly bounded operator valued function $K^N(\cdot) : [t_0, \infty) \mapsto L(H^N, U^N)$ such that

$$\|K^N(t)\|_{L(H^N,U^N)} \leq M_K,$$

and if T_K^N is the evolution operator corresponding to the perturbation of T^N by $B^N K^N(\cdot)$, then

$$\|T_K^N(t,s)\|_{L(H^N)} \leq Me^{-\omega(t-s)}, \quad \text{for } (t,s) \in \triangle_\infty(t_0).$$

Condition UD (*Uniform detectability*). There exist constants $M_\Psi, M, \omega > 0$ independent of N such that for each of the approximation systems, we can find a uniformly bounded operator valued function $\Psi^N(\cdot) : [t_0, \infty) \mapsto L(H^N)$ such that

$$\|\Psi^N(t)\|_{L(H^N)} \leq M_\Psi,$$

and if T_Ψ^N is the evolution operator corresponding to the perturbation of T^N by $\Psi^N(W^N)^{1/2}(\cdot)$, then

$$\|T_\Psi^N(t,s)\|_{L(H^N)} \leq Me^{-\omega(t-s)}, \quad \text{for } (t,s) \in \triangle_\infty(t_0).$$

We may summarize our findings as follows.

THEOREM 4.5. *Under Hypotheses* 4.1–4.3, 4.5–4.7, *the conditions* $\|P_H^N\varphi - \varphi\|_V \to 0$ *for* $\varphi \in V$, $\|P_U^N v - v\|_U \to 0$ *for* $v \in U$, *and the Conditions US and UD, there exists a unique uniformly bounded solution* Q^N *of the Riccati integral equation on the infinite time interval* $[t_0, \infty)$ *for each approximate system in* H^N. *Furthermore, the sequence* $Q^N(t)P_H^N$ *converges strongly to* $Q(t)$ *and the convergence is uniform in* t *for* t *in any bounded interval.*

These results follow from Theorem 3.2 and our discussions above. We have thus reduced our problem of ensuring convergence of the Riccati variables to one of guaranteeing uniform stabilizability and detectability of the approximate systems. In the following two sections, two different approaches to obtaining uniform detectability (Condition UD) and stabilizability (Condition US) are presented.

5. Dissipativity and uniform stabilizability / detectability. The original control system (4.10) defined in §4 was assumed to be stabilizable and detectable (i.e., we assumed Hypotheses 4.5, 4.6 held). For a given evolution system, often an easy way to ascertain stability is using the dissipativity of the system. In particular, if a system satisfies Hypothesis 4.4, by Theorem 4.3 the associated evolution operator is uniformly exponentially stable. This naturally suggests a sufficient condition for stabilizability of a control system.

Hypothesis 5.1. There exists a uniformly bounded function $K(\cdot) : [t_0, \infty) \mapsto L(H, U)$ and a constant $k > 0$ such that for all $\varphi \in V$,

$$\sigma(t; \varphi, \varphi) + < B(t)K(t)\varphi, \varphi >_H \geq k\|\varphi\|_H^2, \quad \text{for } t \in [t_0, \infty).$$

LEMMA 5.1. *Under Hypothesis* 5.1, *the control system defined by* (4.10)–(4.12) *is stabilizable. In fact if* T_K *is the evolution operator corresponding to the perturbation of the evolution operator* T *by* $-BK$, *we can find constants* $M, \alpha > 0$ *such that*

$$\|T_K(t, s)\|_{L(H)} \leq Me^{-\alpha(t-s)}, \quad \textit{for } (t, s) \in \triangle_\infty(t_0).$$

As a consequence, there exists a constant C *such that for all* $x_0 \in H$, *we can find a control* $u(\cdot) \in L^2([s, \infty); U)$ *with a cost*

$$J_\infty(u; x_0, s) \leq C\|x_0\|^2, \quad \textit{for } s \in [t_0, \infty).$$

Proof. Let $K(\cdot)$ be the operator valued function in Hypothesis 5.1. Define the perturbed sesquilinear form $\sigma_K(t; \varphi, \psi) = \sigma(t; \varphi, \psi) + < B(t)K(t)\varphi, \psi >_H$. Then T_K is associated with σ_K as in Theorem 4.2 with $-B(t)K(t)$ as the perturbation term. Under our assumptions, by Theorem 4.3, there exist $M, \alpha > 0$ such that

$$\|T_K(t, s)\|_{L(H)} \leq Me^{-\alpha(t-s)}, \quad \text{for } (t, s) \in \triangle_\infty(t_0).$$

For any $x_0 \in H$, let $v(t) = -K(t)T_K(t, s)x_0$; it is easy to see that the corresponding trajectory is $x(t) = T_K(t, s)x_0$.

By our standing assumptions, the operator valued functions $W(\cdot), R(\cdot), B(\cdot)$ are uniformly bounded in the entire interval $[t_0, \infty)$. We choose C such that

$$C \geq (\|W(t)\|_{L(H)} + \|K^*(t)R(t)K(t)\|_{L(H)})M^2/2\alpha, \quad \text{for } t \geq t_0.$$

Then

$$\begin{aligned} J_\infty(v; x_0, s) &= \int_s^\infty \{< W(t)x(t), x(t) >_H + < R(t)K(t)x(t), K(t)x(t) >_U\} dt \\ &\leq \int_s^\infty \left\{\|W(t)\|_{L(H)} + \|K^*(t)R(t)K(t)\|_{L(H)}\right\} M^2 e^{-2\alpha(t-s)}\|x_0\|_H^2 dt \\ &\leq C\|x_0\|_H^2. \end{aligned}$$

Similarly, a sufficient condition for detectability can be stated as follows.

Hypothesis 5.2. There exists a uniformly bounded operator valued function $\Psi(\cdot) : [t_0, \infty) \mapsto L(H)$, and constant $\lambda > 0$ such that

$$\sigma(t; \varphi, \varphi) + < \Psi(t)W^{\frac{1}{2}}(t)\varphi, \varphi >_H \geq \lambda\|\varphi\|_H^2, \quad \varphi \in V.$$

LEMMA 5.2. *Under Hypothesis* 5.2, *the control system defined by* (4.10)– (4.12) *is detectable.*

In the remainder of the current section, we assume that Hypotheses 5.1, 5.2 hold for our control system in H. The strict H-dissipativity Hypotheses 5.1, 5.2 on the evolution systems are stronger than the usual stabilizability and detectability hypotheses; however, they are in general easy to verify for a wide class of problems. Moreover, the constants M and the decay rates α depend only on the values of k and λ (which, of course, are dependent on K and Ψ). Thus, this type of approach suggests that approximate systems which preserve the H-dissipativity might be uniformly stabilizable and uniformly detectable. Pursuing this type of argument, we shall try to show that the following conditions are implied by Hypothesis 5.1, 5.2. (As in the discussions of §4 surrounding (4.17)– (4.19), we assume that $B^N(t) = P_H^N B(t)$ and $W^N(t) = P_H^N W(t)$.)

Condition 5.1. There exists a constant $\tilde{k} > 0$ such that for every N, there exists a uniformly bounded operator valued function $K^N(\cdot) : [t_0, \infty) \mapsto L(H^N, U^N)$ so that

$$\sigma^N(t; \varphi^N, \varphi^N) + < B^N(t)K^N(t)\varphi^N, \varphi^N >_H \geq \tilde{k}\|\varphi^N\|_H^2 \tag{5.1}$$

holds for all $\varphi^N \in H^N$.

Condition 5.2. There exists a constant $\tilde{\lambda} > 0$ such that for every N, there exists a uniformly bounded operator valued function $\Psi^N(\cdot) : [t_0, \infty) \mapsto L(H^N)$ so that

$$\sigma^N(t; \varphi^N, \varphi^N) + < \Psi^N(t)(W^N(t))^{\frac{1}{2}}\varphi^N, \varphi^N >_H \geq \tilde{\lambda}\|\varphi^N\|_H^2 \tag{5.2}$$

holds for all $\varphi^N \in H^N$.

Note that if the original system satisfies Hypotheses 5.1 and 5.2, by the definition of σ^N we have

$$\sigma^N(t; \varphi^N, \varphi^N) + < P_H^N B(t)K(t)\varphi^N, \varphi^N >_H \geq k\|\varphi^N\|_H^2 \tag{5.3}$$

$$\sigma^N(t; \varphi^N, \varphi^N) + < P_H^N \Psi(t)W^{\frac{1}{2}}(t)\varphi^N, \varphi^N >_H \geq \lambda\|\varphi^N\|_H^2. \tag{5.4}$$

Let us compare inequality (5.3) to (5.1); if we could take $K^N(t) = K(t)$ in (5.1), then Condition 5.1 holds trivially. However, a careful examination of inequalities (5.3) and (5.4) reveals that they do not provide stabilizability and detectability of the approximate system. In the case of (5.3) vs. (5.1) we observe that the range of the operator $K(\cdot)$ hypothesized in Hypothesis 5.1 is not necessarily in U^N, and $K(\cdot)$ cannot be used as a stability operator for the control system in H^N, U^N as required in Condition 5.1. Comparing (5.4) and (5.2) and recalling that $W^N(t) = P_H^N W(t)$, we see that the the choice $\Psi^N = P_H^N \Psi P_H^N$ would suffice only in the case where $P_H^N (P_H^N W(t))^{1/2} = W^{1/2}(t)$.

Before we state additional conditions for the approximate systems, let us consider several interesting cases for which stabilizability and detectability are preserved.

Case I. *Dissipative systems.* Suppose that Hypotheses 5.1, 5.2 hold for $K(t) \equiv 0$ and $\Psi(t) \equiv 0$; then by definition of the sesquilinear form σ^N, Conditions 5.1, 5.2 hold

with $K^N(t) \equiv 0$ and $\Psi^N(t) \equiv 0$. This is the case when the homogeneous system is itself dissipative.

Case II. *Finite-dimensional control.* Suppose the control space U is finite-dimensional. Taking $U^N = U$, we can use $K^N(t) = K(t)$ and the approximate systems are uniformly stabilizable.

Case III. *Special stability operators.* Consider the inequalities in Hypotheses 5.1, 5.2 again. We can assume that these inequalities hold where $B(t)K(t)$ and $\Psi(t)W^{1/2}(t)$ are nonnegative definite self-adjoint operators. Suppose that there exist scalar functions $\kappa(t) \geq 0$ and $\mu(t) \geq 0$ such that

$$B(t)K(t) \leq \kappa(t)B(t)B^*(t), \quad \Psi(t)W^{1/2}(t) \leq \mu(t)W(t).$$

Then taking $K(t) = \kappa(t)B^*(t)$ and $\Psi(t) = \mu(t)W^{1/2}(t)$, we find Hypotheses 5.1 and 5.2 also hold. If we modify slightly the definition of the sesquilinear form σ^N by

$$\begin{aligned}\hat{\sigma}^N(t;\varphi^N,\psi^N) &= \sigma^N(t;\varphi^N,\psi^N) + < B(t)[I - P_U^N]K(t)\varphi^N, \psi^N > \\ &\quad + < \Psi(t)[I - P_H^N]W^{\frac{1}{2}}(t)\varphi^N, \psi^N >,\end{aligned}$$

the sesquilinear form $\hat{\sigma}^N$ satisfies Conditions 5.1 and 5.2. Indeed, we note that the perturbation terms satisfy

$$\begin{aligned}&< B(t)[I - P_U^N]K(t)\varphi^N, \varphi^N > \\ &\quad = \kappa(t) < [I - P_U^N]B^*(t)\varphi^N, [I - P_U^N]B^*(t)\varphi^N >_U \geq 0, \\ &< \Psi(t)[I - P_H^N]W^{1/2}\varphi^N, \varphi^N > \\ &\quad = \mu(t) < [I - P_H^N]W^{1/2}(t)\varphi^N, [I - P_H^N]W^{1/2}(t)\varphi^N >_H \geq 0.\end{aligned}$$

Thus by taking $K^N(t) = P_U^N K(t)$, $\Psi^N(t) = P_H^N \Psi(t)$, Conditions 5.1, 5.2 hold for $\hat{\sigma}^N$. On the other hand, the perturbation terms go to zero as N goes to ∞. Therefore if we use $\hat{\sigma}^N$ as the sesquilinear form for the approximate control system in H^N, the corresponding evolution operator $\hat{T}^N$ should also converge to T.

The three cases above motivate us to consider the following modifications of the sesquilinear form in H^N. Let the operator valued functions $K(\cdot), \Psi(\cdot)$ be as in Hypotheses 5.1 and 5.2; define $\hat{\sigma}^N$ as:

$$\begin{aligned}\hat{\sigma}^N(t;\varphi^N,\psi^N) &= \sigma^N(t;\varphi^N,\psi^N) + < B(t)[I - P_U^N]K(t)\varphi^N, \psi^N > \\ &\quad + < \Psi(t)[I - P_H^N]W^{\frac{1}{2}}(t)\varphi^N, \psi^N >,\end{aligned}$$

for all $\varphi^N, \psi^N \in H^N$. Let $\hat{A}^N(t) : H^N \mapsto H^N$ be defined by

$$- < \hat{A}^N(t)\varphi^N, \psi^N >_H = \hat{\sigma}^N(t;\varphi^N,\psi^N), \quad \varphi^N, \psi^N \in H^N.$$

Let $\hat{T}^N(\cdot,\cdot)$ be the evolution operator generated by $\hat{A}^N(t)$.

We can repeat the arguments in the proof of Theorem 4.4 using $\hat{T}^N$ in place of T^N. In the arguments, there is an extra term

$$2\int_s^t \mathrm{Re}\ < \Lambda^N(\eta)w^N(\eta), z^N(\eta) >_H \, d\eta$$

on the right side of the inequalities, where $\Lambda^N(\eta) \in L(H^N)$ is given by

$$\Lambda^N(t) = P_H^N B(t)[I - P_U^N]K(t) + P_H^N \Psi(t)[I - P_H^N]W^{1/2}(t).$$

There exists a constant C independent of N (the projections P_H^N and P_U^N are convergent) such that

$$\|\Lambda^N(\eta)\|_{L(H^N)} \leq C, \quad \eta \in [a,b],$$

and, furthermore,

$$\|\Lambda^N(\eta)P_H^N\varphi\|_H \to 0, \quad \text{for } \varphi \in H.$$

Recalling $z^N = w - w^N$ and using (4.15), we have $w^N = w - z^N = P_H^N w - P_H^N z^N$. We thus find (suppressing the argument η throughout)

$$\begin{aligned} \text{Re } < \Lambda^N w^N, z^N >_H &= \left| < \Lambda^N (P_H^N w - P_H^N z^N), P_H^N z^N >_H \right| \\ &\leq \|\Lambda^N P_H^N w\|_H \|P_H^N z^N\|_H + C\|P_H^N z^N\|_H^2 \\ &\leq \frac{1}{2}\|\Lambda^N P_H^N w\|_H^2 + (C + \frac{1}{2})\|P_H^N z^N\|_H^2. \end{aligned}$$

The integral (with respect to η) of the first term in this last expression $\to 0$ uniformly in t, s and can be added to the term $\Theta^N(t,s)$ in (4.16), while the integral of the second term can be included with the integral term in the right side of (4.16). We thus have the following argument.

THEOREM 5.1. *Under Hypotheses* 5.1, 5.2, *the conclusions* (i), (ii) *of Theorem* 4.4 *hold for* $\hat{T}^N$.

Now consider $\hat{T}^N$ as the evolution operator for our approximate control systems in H^N; the convergence of the solutions of the Riccati integral equation in any finite time interval still holds. To generalize the arguments in the three special cases I–III above, we make the following additional assumptions.

Hypothesis 5.3. Consider $K(\cdot), \Psi(\cdot)$ as in Hypotheses 5.1, 5.2 and assume there exist constants $\hat{k} < k$, $\hat{\lambda} < \lambda$ and $\hat{N}$ such that for all $N \geq \hat{N}$

$$\begin{aligned} < (I - P_U^N)K(t)\varphi^N, (I - P_U^N)B^*(t)\varphi^N > &\geq -\hat{k}\|\varphi^N\|_H^2, \\ < (I - P_H^N)W^{\frac{1}{2}}(t)\varphi^N, (I - P_H^N)\Psi^*(t)\varphi^N > &\geq -\hat{\lambda}\|\varphi^N\|_H^2, \end{aligned}$$

for all $\varphi^N \in H^N$.

LEMMA 5.3. *Under Hypotheses* 5.1–5.3, *the approximate control systems are uniformly stabilizable and detectable.*

Proof. We assume without loss of generality that $\hat{N} = 1$. Let $\tilde{k} = k - \hat{k}$ and $\tilde{\lambda} = \lambda - \hat{\lambda}$. By Hypothesis 5.3, $\tilde{k} > 0$ and $\tilde{\lambda} > 0$. Take $K^N(t) = P_U^N K(t)$ and $\Psi^N(t) = P_H^N \Psi(t)$; then with this choice of K^N, Ψ^N, Conditions 5.1, 5.2 hold for $\hat{\sigma}^N$. Let $\hat{T}_K^N$, $\hat{T}_\Psi^N$ be evolution operators corresponding to the perturbations of $\hat{T}^N$ by $B^N K^N$ and $\Psi^N (W^N)^{1/2}$, respectively. By Theorem 4.3, there exist constants $M, \alpha > 0$ depending on $\tilde{k}$, $\tilde{\lambda}$ only such that

$$\|\hat{T}_K^N(t,s)\|_{L(H^N)} \leq Me^{-\alpha(t-s)}, \quad \|\hat{T}_\Psi^N(t,s)\|_{L(H^N)} \leq Me^{-\alpha(t-s)}.$$

By the general framework of §3, there exists a unique solution $\hat{Q}^N$ of the Riccati equation on the infinite time interval for each control system in H^N. The operator

$\hat{Q}^N(t)P_H^N$ converges strongly to the unique solution $Q(t)$ of the Riccati integral equation for the original system in H as $N \to \infty$. The convergence is uniform in t for t in any bounded interval.

We summarize the results for our dissipativity approach to uniform stabilizability and detectability in the following theorem.

THEOREM 5.2. *Consider the parabolic control system defined by* (4.10)– (4.12) *under Hypotheses* 4.1–4.3, 4.5–4.7, *and the corresponding approximate systems as defined via* $\hat{\sigma}^N$, $\hat{T}^N$ *as above where* $P_H^N \to I$ *strongly in* V *and* $P_U^N \to I$ *strongly in* U. *Under the "H-dissipativity" Hypotheses* 5.1, 5.2 *and the consistency Hypothesis* 5.3, *the following conclusions hold:*

(i) *There exist unique uniformly bounded solutions* $\hat{Q}^N, Q$ *of the Riccati integral equations in the infinite time interval* $[t_0, \infty)$ *for each of the approximate control systems and the original system, respectively. There exists a constant* M *such that for all* $t \in [t_0, \infty)$ *and all* N

$$\|\hat{Q}^N(t)\|_{L(H^N)} \leq M, \quad \|Q(t)\|_{L(H)} \leq M.$$

(ii) *Let* $\hat{S}^N$, S *be the perturbed evolution operator corresponding to the perturbations of* $\hat{T}^N, T$ *by* $-B^N(R^N)^{-1}B^{*N}\hat{Q}^N$ *and* $-BR^{-1}BQ$, *respectively; then there exist constants* M *and* $\alpha > 0$ *independent of* N *such that*

$$\|\hat{S}^N(t,s)\|_{L(H^N)} \leq Me^{-\alpha(t-s)}, \quad \|S(t,s)\|_{L(H)} \leq Me^{-\alpha(t-s)},$$

for all $(t,s) \in \triangle_\infty(t_0)$.

(iii) *As* $N \to \infty$, $\hat{Q}^N, \hat{S}^N$ *converge to* Q, S *in the following sense: for all* $\varphi \in H$

$$\|\hat{Q}^N(t)P_H^N\varphi - Q(t)\varphi\|_H \to 0,$$
$$\|\hat{S}^N(t,s)P_H^N\varphi - S(t,s)\varphi\|_H \to 0.$$

The convergence is uniform in (t,s) *in any bounded interval.*

The advantage of using the dissipativity approach outlined above is that the hypotheses are readily checked. The H-dissipativity can sometimes be replaced by even weaker dissipativity conditions for which one can obtain an exponential decay rate (e.g., see [Ch], [La]). For parabolic systems with strict V-ellipticity, we can avoid use of this type of argument as we shall see in the next section. However these results might be useful for systems without strict V-ellipticity or possibly even some hyperbolic systems (e.g., see [BKS], [BKW]).

6. Periodic systems: compactness and uniform stabilizability/detectability. One of the special features of parabolic evolution systems as defined in §4 is that the evolution operator $T(t,s)$ is also a bounded linear operator from H to V. Since often the space V is compactly embedded in H, $T(t,s)$ is thus a compact operator. Using this fact, we can show that the convergence of the sequence of operators $T^N(t,s)$ to $T(t,s)$ is in a stronger sense. In this section, by combining periodicity and compactness of the evolution operators T^N and T, we can show that the approximation schemes discussed in §4 preserve detectability and stabilizability.

The fundamental ideas on the stability of periodic evolution operators used in our arguments here can be found in [H1], [H2]. The use of compact embedding ideas for the proof of operator norm convergence can be found in [BI2] (The authors gratefully acknowledge K. Ito for fruitful discussions regarding this approach).

In this section we make a periodicity assumption for our control system:

Hypothesis 6.1. There exists a constant $\theta > 0$ such that
(i) The sesquilinear form σ is θ-periodic in time;
(ii) The operator valued functions B, R, W are periodic in time with period θ.

LEMMA 6.1. *Under the above assumption, the evolution operator $T(\cdot,\cdot)$ of the corresponding homogeneous evolution equation is θ-periodic.*

Proof. For any $s \leq t$, and all $\varphi, \psi \in V$, we have

$$\begin{aligned} < T(t+\theta, s+\theta)\varphi, \psi >_H &= < \varphi, \psi >_H - \int_{s+\theta}^{t+\theta} \sigma(\tau; T(\tau, s+\theta)\varphi, \psi) d\tau \\ &= < \varphi, \psi >_H - \int_s^t \sigma(\tau; T(\tau+\theta, s+\theta)\varphi, \psi) d\tau. \end{aligned}$$

By the uniqueness of the solution of the weak form of our evolution equation, we have $T(t+\theta, s+\theta)\varphi = T(t,s)\varphi$, for $\varphi \in H$.

Under the periodicity Hypothesis 6.1, the continuity assumptions and the uniform boundedness assumptions of the control system need only to be verified in the bounded interval $[0, \theta]$. The above lemma shows that the periodicity of the evolution operator is given by the periodicity of the corresponding sesquilinear form σ. The following theorem plays a very important role in the study of periodic systems. We give its proof in order to remind the reader of the dependency of certain bounds involved.

THEOREM 6.1. ([H1], [H2]) *Let $T(\cdot,\cdot)$ be a θ-periodic evolution operator. Then $T(\cdot,\cdot)$ is uniformly exponentially stable if and only if there exist an integer n and a constant $\beta < 1$ such that*

$$\|T(n\theta, 0)\|_{L(H)} \leq \beta. \tag{6.1}$$

Proof. a) Let $T(\cdot,\cdot)$ be uniformly exponentially stable; that is, there exist constants $M, \omega > 0$ such that

$$\|T(t,s)\|_{L(H)} \leq M e^{-\omega(t-s)}, \quad (t,s) \in \triangle_\infty(0).$$

Therefore, if we take n large enough such that $M \exp\{-\omega n\theta\} < 1$, and let $\beta = M \exp\{-\omega n\theta\}$, we have that (6.1) holds.

b) Suppose (6.1) holds. Let C be a constant such that for all $0 \leq s \leq t \leq n\theta$, $\|T(t,s)\| \leq C$. Now for any $0 \leq s \leq t < \infty$ and $t - s > \theta$, we can find integers k and m such that

$$k \leq \frac{t-s}{n\theta} \leq k+1, \quad (m-1)\theta \leq s \leq m\theta.$$

Therefore, by the semigroup property of $T(\cdot,\cdot)$, we get

$$T(t,s) = T(t, (nk+m)\theta) T((nk+m)\theta, m\theta) T(m\theta, s).$$

By definition of k and m, we have $m\theta - s \leq n\theta$ and $t - (nk+m)\theta \leq n\theta$; then by (6.1), we have

$$\|T(t,s)\|_{L(H)} \leq C^2 \beta^k \leq C^2 e^{k \log \beta}.$$

Since $\beta < 1$, we have $\log \beta < 0$; therefore we find

$$\|T(t,s)\|_{L(H)} \leq C^2 e^{-\log \beta} \cdot e^{(k+1)\log \beta} \leq M e^{-\omega(t-s)}.$$

with $M = C^2/\beta$ and $\omega = -\log \beta / n\theta$.

Since the evolution operator T is also $n\theta$-periodic, we can therefore assume without loss of generality that $\|T(\theta,0)\|_{L(H)} \leq \beta < 1$. The following lemma is an interesting consequence of the above theorem.

LEMMA 6.2. *For a periodic system, if the stabilizability and the detectability assumptions are satisfied, then there exist θ-periodic operator valued functions $\hat{K}(\cdot)$ and $\hat{\Psi}(\cdot)$ such that the evolution operator $\hat{T}_K$ and $\hat{T}_\Psi$ corresponding to the perturbation of the evolution operator T by $B\hat{K}$, $\hat{\Psi}W^{1/2}$ are also uniformly exponentially stable.*

Proof. Suppose K, Ψ are operator valued functions in the stabilizability and detectability assumptions. Let T_K, T_Ψ be the evolution operators corresponding to perturbation of T by BK and $\Psi W^{1/2}$ respectively. Without lost of generality, we can assume that there exists a constant $\beta < 1$ such that

$$\|T_K(\theta,0)\| \leq \beta; \tag{6.2}$$

$$\|T_\Psi(\theta,0)\| \leq \beta. \tag{6.3}$$

Now define θ-periodic operator valued functions $\hat{K}, \hat{\Psi}$ as $\hat{K}(t) = K(t)$, $\hat{\Psi}(t) = \Psi(t)$, for $t \in [0,\theta)$, and extend periodically for $t \geq \theta$. Then we have $\hat{T}_K(\theta,0) = T_K(\theta,0)$, $\hat{T}_\Psi(\theta,0) = T_\Psi(\theta,0)$; therefore (6.2), (6.3) still hold for the new evolution systems. By Theorem 6.1, we conclude that $\hat{T}_K$, $\hat{T}_\Psi$ are also uniformly exponentially stable.

As a consequence of this lemma, we can assume without loss of generality that the operator valued functions K, Ψ in Hypotheses 4.5 and 4.6 are θ-periodic. In fact we can make the following equivalent assumptions.

Hypothesis 6.2. There exist a constant $\beta < 1$ and θ-periodic operator valued functions K, Ψ, such that if T_K, T_Ψ are the evolution operators corresponding to the perturbation of the evolution operator T by $BK, \Psi W^{1/2}$, respectively, then

$$\|T_K(\theta,0)\| \leq \beta, \quad \|T_\Psi(\theta,0)\| \leq \beta.$$

For the remainder of this section, we shall assume Hypotheses 6.1 and 6.2 hold and we focus on the uniform stabilizability and the uniform detectability of the approximate control systems. Let $K^N(t) = P_U^N K(t)$, $\Psi^N(t) = P_H^N \Psi$ and T_K^N, T_Ψ^N be the evolution operators corresponding to the perturbations of T^N by $B^N K^N$ and $\Psi^N (W^N)^{1/2}$, respectively. As we have seen in the proof of the Theorem 6.1, the constant M and decay rate ω depend only on the uniform bound of T and constant β. We already know (use the arguments of Theorem 4.4 and uniform boundedness of the perturbations) that $\|T_K^N(t,s)\|_{L(H^N)}$ and $\|T_\Psi^N(t,s)\|_{L(H^N)}$ can be uniformly bounded by a constant C independent of N for all $0 \leq s \leq t \leq \theta$; therefore, to obtain uniform stabilizability and uniform detectability, we only have to show (see b) of the proof of Theorem 6.1) that we can find $\hat{\beta} < 1$ and N_0 such that for all $N \geq N_0$, we have

$$\|T_K^N(\theta,0)\|_{L(H^N)} \leq \hat{\beta}; \tag{6.4}$$

$$\|T_\Psi^N(\theta,0)\|_{L(H^N)} \leq \hat{\beta}. \tag{6.5}$$

If, on the other hand, the convergence of $T_K^N(t,s)P_H^N$ to $T_K(t,s)$ ($T_\Psi^N(t,s)P_H^N$ to $T_\Psi(t,s)$) is in the operator norm, then we can readily establish that (6.4), (6.5) hold for N large enough. The following theorem is very useful in the proof of this desired convergence.

THEOREM 6.2. *Let H, V be Hilbert spaces as defined in §4 with V compactly embedded in H. Consider a sequence of bounded linear operators $\mathcal{T}^N$ defined on H and bounded linear operator $\mathcal{T}$ defined on H. Suppose the range of $\mathcal{T}^N$ and $\mathcal{T}$ are in V, and the following conditions hold:*

(i) *There exists a constant C such that*

$$\|\mathcal{T}^N\|_{L(H,V)} \le C, \quad \|\mathcal{T}\|_{L(H,V)} \le C;$$

(ii) *For any $\varphi \in H$, we have*

$$\|\mathcal{T}^N\varphi - \mathcal{T}\varphi\|_H \to 0, \quad \|\mathcal{T}^{*N}\varphi - \mathcal{T}^*\varphi\|_H \to 0,$$

as $N \to \infty$.

Then the convergence of the sequence of operators $\mathcal{T}^N$ to $\mathcal{T}$ is in the operator norm, that is, $\|\mathcal{T}^N - \mathcal{T}\|_{L(H)} \to 0$, as $N \to \infty$.

Before we give the proof of this theorem, let us state a useful lemma.

LEMMA 6.3. *Consider a nonincreasing sequence of compact sets $E_k \subset H$, $E_k \supseteq E_{k+1}$, $k = 1, 2, 3, \ldots$. If we have $\bigcap_{k=1}^\infty E_k = \{0\}$, then for each $\epsilon > 0$, we can find k_0 large enough such that for all $k \ge k_0$, E_k is a subset of a ball $B(0,\epsilon)$ in H defined by $B(0,\epsilon) = \{\varphi \in H \mid \|\varphi\|_H \le \epsilon\}$.*

Proof. Suppose there exists $\epsilon > 0$ such that for every k, we can find $\varphi_k \in E_k$, such that $\|\varphi_k\|_H > \epsilon$. Since the sequence $\{\varphi_k\}$ is in E_1 which is compact, we can assume that φ_k converges to an element φ in E_1. Obviously $\|\varphi\|_H > \epsilon/2$. On the other hand, φ must be in E_k for all k, therefore φ is also in the set $E = \bigcap_{k=1}^\infty E_k$. But since $E = \{0\}$, this is a contradiction.

Proof of Theorem 6.2. i) By definition of the operator norm, we have

$$\|\mathcal{T}^N - \mathcal{T}\|_{L(H)} = \sup\left\{\|\mathcal{T}^N\varphi - \mathcal{T}\varphi\|_H \mid \varphi \in H, \|\varphi\|_H \le 1\right\}.$$

Now let us define the set F_k as

$$F_k = \bigcup_{N=k}^\infty \{\mathcal{T}^N\varphi - \mathcal{T}\varphi \mid \varphi \in H, \|\varphi\|_H \le 1\}.$$

Let E_k be the H closure of F_k. The sequence of operators $\mathcal{T}^N$ converges to $\mathcal{T}$ in operator norm if and only if for all $\epsilon > 0$, we can find k_0 such that for all $k \ge k_0$, we have $E_k \subseteq B(0,\epsilon)$.

ii) We observe that E_k is a closed set in H and hence in V, and by our assumption $E_k \subset V$ is bounded in V norm, in fact

$$E_k \subset \{\varphi \mid \varphi \in V, \|\varphi\|_V \le 2C\}.$$

Therefore, E_k is a compact set. By definition of E_k, we know that $\{E_k\}$ is a nonincreasing sequence of compact sets.

iii) Let us define $E = \bigcap_{k=1}^{\infty} E_k$. Suppose $\varphi \in E$; since φ is in the closure of F_k for all k, we can find a sequence γ^N with $\|\gamma^N\|_H \leq 1$, such that $\varphi^N = (\mathcal{T}^N - \mathcal{T})\gamma^N$ converges to φ. Therefore, for any $\psi \in H$, we have

$$\begin{aligned} < \varphi, \psi >_H &= \lim_{N\to\infty} < \varphi^N, \psi >_H \\ &= \lim_{N\to\infty} < \gamma^N, (\mathcal{T}^{*N} - \mathcal{T}^*)\psi >_H . \end{aligned}$$

Since γ^N is uniformly bounded in N and $(\mathcal{T}^{*N} - \mathcal{T}^*)\psi$ goes to zero as $N \to \infty$, we have $< \varphi, \psi >_H = 0$ for all $\psi \in H$, and therefore $\varphi = 0$.

Using the previous lemma and i), we obtain $\|\mathcal{T}^N - \mathcal{T}\|_{L(H)} \to 0$, as $N \to \infty$.

Now recall the definition of the approximate control systems defined in §4, and consider $\mathcal{T}^N = T_K^N(\theta, 0)P_H^N$, $\mathcal{T} = T_K(\theta, 0)$. By the results in §4, we can easily verify that the assumptions of Theorem 6.2 are satisfied. Therefore, we have

$$\|T_K^N(\theta, 0)P_H^N - T(\theta, 0)\|_{L(H)} \to 0,$$

as $N \to \infty$. Now let β be the constant in Hypothesis 6.2; letting $\epsilon = 1 - \beta$, we can find N_0 large enough such that for all $N \geq N_0$ we have

$$\|T_K^N(\theta, 0)P_H^N - T(\theta, 0)\|_{L(H)} \leq \frac{\epsilon}{2}.$$

Therefore, for all $N \geq N_0$, we have

$$\begin{aligned} \|T_K^N(\theta, 0)\|_{L(H^N)} &\leq \|T_K(\theta, 0)\|_{L(H)} + \|T_K^N(\theta, 0)P_H^N - T_K(\theta, 0)\|_{L(H)} \\ &\leq 1 - \frac{\epsilon}{2}. \end{aligned}$$

We summarize our discussion in the following theorem.

THEOREM 6.3. *Let H, V be the Hilbert spaces used in the §4 and assume that V is compactly embedded in H. Suppose that Hypotheses 4.1–4.3, 4.7, and 6.1, 6.2 hold. Let $H^N \subset V$ be the finite dimensional approximation spaces and let the approximate control system be defined as in §4. Then there exists N_0 large enough such that for all $N \geq N_0$, the approximate control systems are uniformly stabilizable and uniformly detectable. As a consequence, if Q^N, Q are the unique solutions of the Riccati integral equations on the infinite time interval in H^N, H, respectively, and S^N, S are the evolution operators corresponding to the perturbations of T^N, T by $-B^N(R^N)^{-1}B^{*N}Q^N$, $-BR^{-1}B^*Q$, respectively, then we have:*

$$\|Q^N(t)P_H^N\varphi - Q(t)\varphi\|_H \to 0, \quad \|S^N(t, s)P_H^N - S(t, s)\|_{L(H)} \to 0,$$

as $N \to \infty$. The convergences are uniform in $[0, \theta]$ and for $(t, s) \in \triangle(0, \theta)$, respectively.

We remark that an autonomous system is a particular case of a periodic system; therefore, the approach used here can also be applied to time invariant systems as considered in [BK1]. In the case of parabolic systems with strict V-ellipticity where V is compactly embedded in H, the arguments in this section offer an alternative (and more succinct) approach to uniform stabilizability/detectability from the dissipativity approach of §5.

7. Parabolic partial differential equation control systems: An example. In this section we consider control systems governed by second-order parabolic partial differential equations with distributed scalar control. We indicate briefly how one formulates the associated control and approximate problems in the framework of §4. For this class of systems we show that one can, under standard assumptions, readily verify the conditions for continuity, ellipticity, stabilizability and detectability required in §4. For Galerkin-type approximation schemes based on spline subspaces, Conditions US and UD are readily established.

Let Ω be a bounded open subset of $\Re^n$ with an infinitely differentiable boundary Γ given by a variety of dimension $n-1$ and consider the following homogeneous second-order parabolic partial differential equation ([L1], [L2, p. 100]):

$$\frac{\partial}{\partial t}z(t,\xi) = \sum_{i,j=1}^{n}\frac{\partial}{\partial \xi_i}\left(a_{i,j}(t,\xi)\frac{\partial}{\partial \xi_j}z(t,\xi)\right)+\sum_{i=1}^{n}b_i(t,\xi)\frac{\partial}{\partial \xi_i}z(t,\xi) + c(t,\xi)z(t,\xi), \quad t>0, \xi\in\Omega, \tag{7.1}$$

where $\xi=(\xi_1,\cdots,\xi_n)\in\Re^n$. Generic boundary conditions are given by

$$\gamma(\xi)\sum_{i,j=1}^{n}a_{i,j}(t,\xi)\frac{\partial}{\partial \xi_j}z(t,\xi)\eta_i(\xi)+\beta(t,\xi)z(t,\xi)=0, \quad t>0,\xi\in\Gamma, \tag{7.2}$$

where $\eta(\xi)=(\eta_1(\xi),\cdots,\eta_n(\xi))$ is the outward unit normal vector at a point ξ on the boundary Γ of Ω. Note that for all $\xi\in\Gamma$, if $\gamma(\xi)\neq 0$, we can divide (7.2) by $\gamma(\xi)$; therefore, we can assume without loss of generality that γ takes only values 0 or 1. We choose as our state space $H=L^2(\Omega)$; the appropriate choice for V depends on the boundary conditions and we consider several special cases.

(i) Consider the case $\gamma(\xi)\equiv 0$, $\beta(t,\xi)\equiv 1$, and thus equation (7.2) specifies the usual Dirichlet boundary condition. We then define $V=H_0^1(\Omega)$, and a sesquilinear form σ_1 by

$$\sigma_1(t;\varphi,\psi) = \int_\Omega \sum_{i,j=1}^{n}a_{i,j}(t,\xi)\frac{\partial}{\partial\xi_j}\varphi(\xi)\frac{\partial}{\partial\xi_i}\bar{\psi}(\xi)d\xi - \int_\Omega\sum_{i=1}^{n}\left\{b_i(t,\xi)\frac{\partial}{\partial\xi_i}\varphi(\xi)\bar{\psi}(\xi)+c(t,\xi)\varphi(\xi)\bar{\psi}(\xi)\right\}d\xi.$$

(ii) If we consider the case for $\gamma(\xi)\equiv 1$, $\beta(t,\xi)\equiv 0$, we obtain a Neumann boundary condition. We then choose $V=H^1(\Omega)$ and note that the integrals in the definition of σ_1 above are also defined for any functions φ,ψ in $H^1(\Omega)$. Therefore, the sesquilinear form σ_2 for Neumann boundary conditions can be taken as the same as for Dirichlet conditions,$\sigma_2=\sigma_1$, and thus only the spaces V are changed.

(iii) Consider the case $\gamma(\xi)\equiv 1$ and $\beta(t,\xi)\not\equiv 0$ which results in Robin or mixed boundary conditions. We again choose $V=H^1(\Omega)$, and define a sesquilinear form σ_3 by

$$\sigma_3(t;\varphi,\psi)=\sigma_2(t;\varphi,\psi)+\int_\Gamma\beta(t,\xi)\varphi(\xi)\bar{\psi}(\xi)d\xi.$$

By taking $\gamma(\xi)=0$ on a part Γ_1 of the boundary and $\gamma(\xi)=1$ on $\Gamma-\Gamma_1$, we can obtain other mixed boundary conditions. The choice of space V should also be

modified accordingly. In all the cases above, let V be the appropriate choice of Sobolev space (either H_0^1 or H^1), and let σ denote the corresponding sesquilinear form defined on $V \times V$. Then a solution $z(t)$ of (7.1) and (7.2) satisfies

$$\frac{d}{dt} < z(t), \psi >_H = -\sigma(t; z(t), \psi), \quad \text{for all } \psi \in V. \tag{7.3}$$

As usual, (7.3) is called the weak form of (7.1) and (7.2); see (4.7), (4.8), and (4.10), (4.11) of §4.

The continuity and the ellipticity conditions for the sesquilinear forms σ_i can be characterized by properties of the coefficients $a_{i,j}, b_i, c$, and β. The standard assumptions ([L2, p.100]) for V-continuity and Hölder continuity of the sesquilinear forms σ_i are as follows: For each fixed t, the functions $a_{i,j}(t,\cdot)$, $b_i(t,\cdot)$, $c(t,\cdot)$ are elements of $L^\infty(\Omega)$, while the function $\beta(t,\cdot)$ is an element of $L^\infty(\Gamma)$. Furthermore, for each bounded interval $[a,b]$, there exist constants $C > 0$ and $0 < \gamma \leq 1$, such that each of the coefficients $a_{i,j}, b_i, c, \beta$ satisfy the bounds $\|f(t,\cdot)\|_{L^\infty} \leq C$ and $\|f(t,\cdot) - f(s,\cdot)\|_{L^\infty} \leq C|t-s|^\alpha$ for t, s in $[a,b]$, where L^∞ is $L^\infty(\Omega)$ or $L^\infty(\Gamma)$ as appropriate.

It is easily seen that under these assumptions, the V-continuity and V-Hölder continuity Hypotheses 4.1 and 4.3 hold for σ_1 and σ_2 defined above. In the case of σ_3, a boundary integral is involved; but under our assumptions on the smoothness of the boundary, the following estimates hold (see [L2, p.17, Thm. 3.2, p.23]): For any $\varphi, \psi \in H^1(\Omega)$, the restrictions of φ, ψ to the boundary Γ belong to $L^2(\Gamma)$. Furthermore, there exists a constant C such that

$$\left| \int_\Gamma \varphi \bar{\psi} d\xi \right| \leq \|\varphi\|_{L^2(\Gamma)} \|\psi\|_{L^2(\Gamma)} \leq C \|\varphi\|_{H^1(\Omega)} \|\psi\|_{H^1(\Omega)}.$$

Furthermore, for all $\epsilon > 0$, there exists constant $C(\epsilon)$ such that

$$\|\varphi\|^2_{L^2(\Gamma)} \leq \epsilon \|\varphi\|^2_{H^1(\Omega)} + C(\epsilon) \|\varphi\|^2_{L^2(\Omega)}.$$

With these estimates, it is readily argued that σ_3 also satisfies the Hypotheses 4.1 and 4.3 of §4.

To assure V-ellipticity we again make standard assumptions: for any bounded interval $[a,b]$, there exists a constant $\nu > 0$ such that for all $t \in [a,b]$ and $\xi \in \Omega$,

$$\nu \sum_{i=1}^n \zeta_i^2 \leq \sum_{i,j=1}^n a_{i,j}(t,\xi) \zeta_i \zeta_j,$$

for all $\zeta = (\zeta_1, \cdots, \zeta_n) \in \Re^n$. Under this assumption, it is readily seen that for each bounded time interval $[a,b]$, there exist constants $c_2 > 0$ and m, such that

$$\text{Re } \sigma_i(t; \varphi, \varphi) \geq c_2 \|\varphi\|^2_V - m \|\varphi\|^2_H,$$

for all $\varphi \in V$ and $t \in [a,b]$. That is, each of the sesquilinear forms σ_i, $i = 1,2,3$, defined above satisfies Hypothesis 4.2.

In the remaining part of the this section, let H, V be the spaces of functions appropriate for a specific problem, and let σ be the sesquilinear form defined for that problem as above. For the control space U, we choose $U = L^2(\Omega)$, with the control system being defined by (see (4.10)–(4.12))

$$\frac{d}{dt} < z(t), \psi >_H = -\sigma(t; z(t), \psi) + < B(t)u(t), \psi >_H, \tag{7.4}$$

for all ψ in V. We choose the cost functional given by (4.12). Here we define the operators $B(t), W(t), R(t)$ by the following

$$\begin{aligned}
[B(t)v](\xi) &= b(t,\xi)v(\xi), \quad \text{for } v \in L^2(\Omega),\\
[W(t)\varphi](\xi) &= w(t,\xi)\varphi(\xi), \quad \text{for } \varphi \in L^2(\Omega),\\
[R(t)v](\xi) &= r(t,\xi)v(\xi), \quad \text{for } v \in L^2(\Omega),
\end{aligned}$$

where b, w, r are scalar valued functions on $[t_0, \infty) \times \Omega$. In this case, the uniform boundedness and the positivity of the corresponding operators can be readily characterized by conditions on the functions b, w, and r. We assume that for each fixed t, $b(t,\cdot), w(t,\cdot)$, and $r(t,\cdot)$ are elements of $L^\infty(\Omega)$. As $L^\infty(\Omega)$ valued functions of t, these functions are assumed continuous. Furthermore, assume there exists a constant C such that for all $t \in [t_0, \infty)$, the functions $b(t,\cdot), w(t,\cdot), r(t,\cdot)$ satisfy the bound $\|f(t,\cdot)\|_{L^\infty(\Omega)} \leq C$. The functions w, r are assumed to be nonnegative and, indeed, we assume there exists a constant $r_0 > 0$ such that $r(t,\xi) \geq r_0$, almost everywhere in Ω, for $t \geq t_0$.

Under these assumptions, the operator valued functions B, W, R satisfy the standing assumptions of §§2 and 4. It remains to consider Hypotheses 4.5 and 4.6 (stabilizability and detectability of the original system) as well as Conditions US and UD once we have introduced approximations. For the problems considered here, we can use the definition of the sesquilinear forms to give sufficient conditions for Hypotheses 5.1 and 5.2 (and hence Hypotheses 4.5 and 4.6) to hold. To this end, we assume that there exist constants $\mu > 0$ and $\rho > 0$ such that for $t \in [t_0, \infty)$, $|b(t,\xi)| \geq \mu$, $|w(t,\xi)| \geq \rho$, almost everywhere in Ω.

Under this assumption, it is readily seen that there exists constants $l \geq 0$ and $k > 0$, such that each of the sesquilinear forms σ_i satisfies for $\varphi \in V$

$$\sigma(t;\varphi,\varphi) + l < B(t)B^*(t)\varphi, \varphi >_H \geq k\|\varphi\|_H^2,$$

and

$$\sigma(t;\varphi,\varphi) + l < W(t)\varphi, \varphi >_H \geq k\|\varphi\|_H^2.$$

Thus, Hypotheses 5.1 and 5.2 hold with $K(t) = lB^*(t)$ and $\Psi(t) = lW^{1/2}(t)$.

For our approximate systems, we choose approximation spaces H^N and U^N as in §§3 and 4 generated by finite element or spline basis elements chosen so that $H^N \subset V$ and $U^N \subset U$ yield the desired convergence properties for P_H^N and P_U^N respectively (see [C, Chaps. 2, 3], [B], [Sc]). The approximating systems are then defined as in §5. It follows immediately that Hypothesis 5.3 holds and hence the conclusions of Theorem 5.2 are valid for the class of examples considered in this section.

We note that under periodicity assumptions we could have applied the alternative approach of §6 to these examples since (see [A]) both $V = H^1(\Omega)$ and $V = H_0^1(\Omega)$ embed compactly in $H = L^2(\Omega)$.

In some of our related efforts, we have numerically tested the ideas presented in this paper on one-dimensional versions of the example of this section. In these examples $\Omega = (0,1)$ and we have to date used either linear or cubic B-splines to generate the approximation spaces H^N and U^N. (In fact, when Ω is a parallelepiped, the above theory still is applicable and tensor products of one dimensional elements are a good choice for approximation elements.) We have considered several examples with time dependent periodic coefficients; for these examples we could use eigendirection

analysis (see [W]) to give an analytic analysis for the feedback control problems. The resulting analytical solutions were used for comparison with the numerical solutions obtained using software implementations based on the theory developed in this paper. Quite satisfactory results were obtained and, as noted in the Introduction, these are being detailed in a separate manuscript under preparation.

In conclusion we note that the theory in this paper is also applicable to higher order parabolic systems (as well as to some boundary damped hyperbolic systems [BKS], [BKW]). In particular, one-dimensional Euler–Bernoulli beam models with Kelvin–Voigt damping satisfy (see [BI1]) the strong ellipticity assumptions needed in the theory developed above. While boundary control (as treated in [BI2]) for such models constitute an obvious class of problems, distributed control as treated in this paper is essential in cases where nonuniform piezoelectric layers along the beam are used to implement the feedback controls.

REFERENCES

[A] R.A. ADAMS, *Sobolev Spaces,* Academic Press, New York, 1975.

[B] C. DE BOOR, *A Practical Guide to Splines,* Springer-Verlag, New York, Berlin, 1978.

[BK1] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems,* SIAM J. Control Optim., 22 (1984), pp. 684 – 698.

[BK2] ——, *Estimation Techniques for Distributed Parameter Systems,* Birkhäuser Boston, 1989.

[BKS] H.T. BANKS, S.L. KEELING, AND R.J. SILCOX, *Optimal control techniques for active noise suppression,* in Proc. 27th IEEE Conf. on Dec. and Control, Austin, TX, Dec. 7–9,1988, pp. 2006–2011.

[BKW] H.T. BANKS, S.L. KEELING, AND C. WANG, *Linear quadratic tracking problems in infinite dimensional Hilbert spaces and a finite dimensional approximation framework,* LCDS/CCS Report #88-28, Brown University, Providence, R.I., 1988.

[BI1] H. T. BANKS AND K. ITO, *A unified framework for approximation in inverse problems for distributed parameter systems,* LCDS/CCS #87-42; Control-Theory and Adv. Tech., 4 (1988), pp. 73–90.

[BI2] H. T. BANKS AND K. ITO, *On a variational approach to a class of boundary control problems: Numerical approximations,* to appear.

[C] P. G. CIARLET, *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, The Netherlands, 1978.

[Ch] G. CHEN, *Energy decay estimates and exact boundary value controllability for wave equation in a bounded domain,* J. Math. Pures Appl., 58 (1979), pp. 249–725.

[CP] 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.

[Da] G. DA PRATO, *Synthesis of optimal control for an infinite dimensional periodic problem,* SIAM J. Control Optim., 25 (1987), pp. 706 –714.

[DI1] G. DA PRATO AND A. ICHIKAWA, *Quadratic control for linear time varying systems,* SIAM J. Control Optim., to appear.

[DI2] G. DA PRATO AND A. ICHIKAWA, *Quadratic control for linear periodic systems,* Appl. Math. Optim., 18 (1988), pp. 39–66.

[Dt] R. DATKO, *Uniform asymptotic stability of evolution processes in a Banach space,* SIAM J. Math. Anal., 3 (1972), pp. 428 – 445.

[FM] H. FUJITA AND A. MIZUTANI, *On the finite element method for parabolic equations,I; Approximation of holomorphic semi-groups,* J. Math. Soc. Japan, 28 (1976), pp. 749 771.

[G] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces,* SIAM J. Control Optim., 17 (1979), pp. 537 – 565.

[H1] J. HALE, *Ordinary Differential Equations,* Second Edition, Robert E. Krieger Publishing Company, Melbourne, FL, 1980.

[H2] J. HALE, *Functional Differential Equations,* Springer-Verlag, New York, Berlin, 1977.

[K] S. G. KREIN, *Linear Differential Equations in Banach Space,* Transl. Math. Monographs, Vol. 29, American Mathematical Society, Providence, R.I., 1971.

[L1] J.L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dériées partielles,* Dunod Gauther-Villars, Paris, 1968.

[L2] ——, *Equations différentelles opérationelles,* Springer-Verlag, New York, Berlin, 1961.

[La] J. LAGNESE, *Decay of solution of wave equations in a bounded region with boundary dissipation,* J. Diff. Equations, 50 (1983), pp. 163–182.

[LT] I. LASIECKA AND R. TRIGGIANI, *The regulator problem for parabolic equations with Dirichlet boundary control,* Parts I, II, Appl. Math. Optim., 16 (1987), pp. 147–168, pp. 187–216.

[R] D.L. RUSSELL, *Mathematics of Finite-Dimensional Control Systems,* Marcel Dekker, New York, 1979.

[S] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations,* Pitman, Boston, 1977.

[Sc] L. SCHUMAKER, *Spline Functions: Basic Theory,* John Wiley & Sons, New York, 1981.

[T] H. TANABE, *Equations of Evolution,* Pitman, Boston, 1979.

[W] C. WANG, *Approximation methods for linear quadratic regulator problems with nonautonomous periodic parabolic systems,* Ph.D Thesis, Brown University, Providence, R.I., May 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1221-1259, November 1989

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.

OPTIMIZATION PROBLEMS IN THE THEORY OF CONTINUOUS TRADING*

IOANNIS KARATZAS†

This paper is dedicated to Dr. Václav E. Beneš on the occasion of his 60th birthday.

Abstract. A unified approach, based on stochastic analysis, to the problems of *option pricing*, *consumption/investment*, and *equilibrium* in a financial market with asset prices modelled by continuous semimartingales is presented.

For the first of these problems, the valuation of both "European" and "American" contingent claims is discussed; the former can be exercised only at a specified time T (the maturity date), whereas the latter can be exercised at any time in $[0, T]$. Notions and results from the theory of *optimal stopping* are employed in the treatment of American options.

A general *consumption/investment* problem is also considered, for an agent whose actions cannot affect the market prices and whose intention is to maximize total expected utility of both consumption and terminal wealth. Under very general conditions on the utility functions of the agent, it is shown how to approach the above problem by considering separately the two, more elementary ones of maximizing utility from consumption only and of maximizing utility from terminal wealth only, and then appropriately composing them. The optimal consumption and wealth processes are obtained quite explicitly. In the case of a market model with constant coefficients, the optimal portfolio and consumption rules are derived very explicitly in feedback form (on the current level of wealth). The Hamilton-Jacobi-Bellman equation of dynamic programming associated with this problem is reduced to the study of two *linear* parabolic equations that are then solved in closed form.

The results of this analysis lead to an explicit computation of the portfolio that maximizes capital growth rate from investment, and to a precise expression for the maximal growth rate.

Finally, the results on the consumption/investment problem for a single agent are applied to study the question of *equilibrium* in an economy with several financial agents whose joint optimal actions determine the price of a traded commodity by "clearing" the markets.

Some familiarity with stochastic analysis, including the fundamental martingale representation and Girsanov theorems, is assumed. Previous exposure to financial economics and/or stochastic control theory is desirable, but not necessary.

Key words. option pricing, consumption/investment optimization, equilibrium, stochastic analysis and control

AMS(MOS) subject classifications. primary 93E20, 90A09; secondary 60G44, 90A16, 49B60, 60G40, 90A14

1. Introduction and summary.

Our aim in this article is to report, hopefully to a wider audience than the already well-informed, on certain recent advances in the theory

* Received by the editors July 12, 1988; accepted for publication (in revised form) January 22, 1989. This research was supported in part by National Science Foundation grant DMS-87-23078.

† Department of Statistics, Columbia University, New York, New York 10027.

of continuous trading which have been made possible thanks to the methodologies of stochastic analysis. All the questions treated here are formulated in the context of a financial market which includes a risk-free asset called the *bond*, and several risky assets called *stocks*; the prices of these latter are driven by an equal number of independent Brownian motions, which model the exogenous forces of uncertainty that influence the market. The interest rate of the bond, the appreciation rates of the stocks as well as their volatilities, constitute the *coefficients* of the market model; we allow them to be arbitrary bounded measurable processes, adapted to the Brownian filtration, but require that a certain nondegeneracy (or "completeness") condition (2.3) be satisfied.

The questions that we address include the following:

(i) A general treatment of the *pricing of contingent claims* such as options, both European (to be exercised only at maturity) and American (which can be exercised any time before or at maturity);

(ii) The resolution of *consumption/investment problems* for a "small investor" (i.e., an economic agent whose actions cannot influence the market prices) with quite general utility functions; and

(iii) The associated study of *equilibrium models*. These are formulated in the context of an economy with several small investors and one commodity, whose price is determined by the joint optimal actions of all these agents in a way that "clears" the markets (i.e., equates supply and demand for the commodity at all times).

Instrumental in the approach that we adopt are two fundamental results of stochastic analysis: the *Girsanov change of probability measure* and the *representation of Brownian martingales as stochastic integrals*. The former constructs processes that are independent Brownian motions under a new, equivalent probability measure which, roughly speaking, "equates the appreciation rates of all the stocks to the interest rate of the bond." The latter of these results provides the "right portfolios" (investment strategies) for the investors in the above-mentioned problems. We assume that the reader is familiar with both these results; they are discussed in several monographs and texts dealing with stochastic analysis, such as Ikeda and Watanabe (1981) and Karatzas and Shreve (1987).

Here is an outline of the paper. Sections 2 and 3 set up the model for the financial market and for the small investor, respectively; the latter has at his disposal the choice of a *portfolio* (investment strategy) and a *consumption strategy*, which determine the evolution of his wealth. The notion of admissible portfolio/consumption strategies, which avoid negative terminal wealth with probability one, is introduced and expounded on in § 4, which can be regarded as the cornerstone of the paper.

Based on the results of § 4, we treat the pricing of European contingent claims in § 5; we provide the fair price and the subsequent values for such instruments, and derive the celebrated Black and Scholes (1973) formula for European call options as a special case of these results. The analogous problems for American contingent claims are taken up in § 6; predictably, their treatment requires notions and results from the theory of optimal stopping.

Sections 7–11 are concerned with *optimization problems* for a small investor. We introduce the concept of utility function in § 7, and treat first a problem in which utility is derived only from consumption (§ 8); based on the methodology of § 4, we provide quite explicit expressions for the optimal consumption and wealth processes, as well as for the associated value $V_1(x)$ of this problem, as a function of the initial wealth $x>0$. The "dual" situation, with utility derived only from terminal wealth, is discussed in § 9; again, explicit expressions are obtained for the above-mentioned quantities,

including the new value function $V_2(x)$. As a byproduct of this analysis, we obtain an explicit computation of the portfolio that maximizes the growth rate from investment, and an equally explicit expression for the maximal capital growth rate (§ 9.6).

We combine the two problems in § 10, where we take up the more realistic case of utility coming from *both* consumption and terminal wealth; it is shown then that a reasonable "compromise" between the two competing objectives can be achieved in the following fashion. At time $t=0$, we let the investor divide his initial capital $x>0$ into two parts $x_1>0$, $x_2>0$, $x_1+x_2=x$; for the initial capital x_1 (respectively, x_2) *he* faces, from then on, a problem in which utility comes only from consumption (respectively, only from terminal wealth). It is shown that this simple procedure, in the form of the superposition of his actions for the two individual problems, yields optimal strategies for the composite problem, provided that x_1 and x_2 are chosen so that $V_1'(x_1)=V_2'(x_2)$. Again, explicit expressions are provided for the optimal consumption and wealth processes, and for the resulting value function $V(x)=V_1(x_1)+V_2(x_2)$.

This type of analysis provides in general no information about the optimal portfolio strategy for the problem of § 10, except for guaranteeing its existence. To amend this drawback, in § 11, we specialize the problem to the case of *constant coefficients*, and reduce the associated Hamilton-Jacobi-Bellman equation to a system of two *linear* parabolic partial differential equations. With the help of the Feynman-Kac theorem and the Black and Scholes formula, we obtain the solutions of these equations in closed form, and from them the value function $V(x)$ by composition; we also derive very explicit expressions for the optimal portfolio and investment strategies, in feedback form on the current level of wealth.

In § 12 we apply the theory of § 8 to the study of an *equilibrium model.* We consider an economy with several agents, who can invest in the financial market and receive continuously endowment streams in units of a certain commodity (consumption good); this latter is traded in the market at a spot price $\psi(\cdot)$. It is shown that the optimal actions of these small investors determine, in principle, the price ψ according to the law of "supply and demand," which mandates that the commodity be consumed entirely as it enters the economy and that the net demand for each financial asset be zero.

2. The financial market model. We shall deal exclusively in this paper with a financial market in which $d+1$ assets (or "securities") can be traded continuously. One of them is a non-risky asset, called the *bond*, with price $P_0(t)$ given by

$$dP_0(t)=P_0(t)r(t)\,dt, \qquad P_0(0)=1. \tag{2.1}$$

The remaining d assets are risky; we shall refer to them as *stocks*, and assume that the price $P_i(t)$ per share of the ith stock is governed by the linear stochastic differential equation

$$dP_i(t)=P_i(t)\left[b_i(t)\,dt+\sum_{j=1}^{d}\sigma_{ij}(t)\,dW_j(t)\right], \qquad P_i(0)=p_i, \quad i=1,2,\cdots,d. \tag{2.2}$$

In this model, $W(t)=(W_1(t),\cdots,W_d(t))^*$ is a Brownian motion in $\mathcal{R}^d$, whose components represent the external, independent sources of uncertainty in the market; with this interpretation, the volatility coefficient $\sigma_{ij}(\cdot)$ in (2.2) models the instantaneous intensity with which the jth source of uncertainty influences the price of the ith stock. Notice that in this model there are as many stocks as there are sources of uncertainty.

The probabilistic setting will be as follows: the Brownian motion W will be defined on the complete probability space $(\Omega, \mathcal{F}, \mathbf{P})$, and we shall denote by $\{\mathcal{F}_t\}$ the $\mathbf{P}$-augmentation of the natural filtration

$$\mathcal{F}_t^W=\sigma(W(s);\,0\leqq s\leqq t), \qquad 0\leqq t<\infty.$$

The *interest rate* process $r(t), 0 \leq t < \infty$ of the bond, the *appreciation rate* vector process $b(t) = (b_1(t), \cdots, b_d(t))^*$, $0 \leq t < \infty$ of the stocks, and the *volatility* matrix-valued process $\sigma(t) = \{\sigma_{ij}(t)\}_{1 \leq i,j \leq d}, 0 \leq t < \infty$, will all be progressively measurable with respect to $\{\mathcal{F}_t\}$, and bounded uniformly in $(t, \omega) \in [0, \infty) \times \Omega$. We shall also assume that the *covariance matrix* process $a(t) = \sigma(t)\sigma^*(t)$ is strongly nondegenerate: i.e., there exists a number $\varepsilon > 0$ such that

$$\xi^* \sigma(t, \omega)\sigma^*(t, \omega)\xi \geq \varepsilon \|\xi\|^2 \quad \forall\, \xi \in \mathcal{R}^d, \quad (t, \omega) \in [0, \infty) \times \Omega. \tag{2.3}$$

We shall refer to $r(\cdot), b(\cdot), \sigma(\cdot)$ collectively as the *coefficients of the market model.*

It follows easily from (2.3) that the inverses of both matrices $\sigma(t, \omega)$ and $\sigma^*(t, \omega)$ exist and are bounded, i.e.,

$$\|(\sigma(t, \omega))^{-1}\xi\| \leq \frac{1}{\sqrt{\varepsilon}} \|\xi\| \quad \forall\, \xi \in \mathcal{R}^d, \tag{2.4}$$

$$\|(\sigma^*(t, \omega))^{-1}\xi\| \leq \frac{1}{\sqrt{\varepsilon}} \|\xi\| \quad \forall\, \xi \in \mathcal{R}^d \tag{2.5}$$

for every $(t, \omega) \in [0, \infty) \times \Omega$.

The nondegeneracy condition (2.3) will allow us to introduce an auxiliary probability measure $\tilde{\mathbf{P}}$, equivalent to $\mathbf{P}$, which is going to be the "catalyst" for all future developments in this paper. To this end, let us introduce the $\mathcal{R}^d$-valued process

$$\theta(t) = (\sigma(t))^{-1}[b(t) - r(t)\mathbf{1}] \tag{2.6}$$

which is bounded, measurable, and adapted to $\{\mathcal{F}_t\}$ thanks to our assumptions, as well as the exponential martingale

$$Z(t) = \exp\left\{ -\int_0^t \theta^*(s)\, dW(s) - \frac{1}{2}\int_0^t \|\theta(s)\|^2\, ds \right\}. \tag{2.7}$$

We fix, from now on, a finite time-horizon $[0, T]$, on which we are going to treat almost all our problems.[1] The auxiliary probability measure $\tilde{\mathbf{P}}$ is defined then on $(\Omega, \mathcal{F}_T)$ by

$$\tilde{\mathbf{P}}(A) = \mathbf{E}[Z(T).1_A], \tag{2.8}$$

and according to the Girsanov theorem the process

$$\tilde{W}(t) = W(t) + \int_0^t \theta(s)\, ds, \qquad 0 \leq t \leq T \tag{2.9}$$

is an $\mathcal{R}^d$-valued Brownian motion under $\tilde{\mathbf{P}}$ (cf. Girsanov (1960) or Karatzas and Shreve (1987, § 3.5)).

In order to understand the significance of the auxiliary probability measure $\tilde{\mathbf{P}}$, let us rewrite (2.2) with the help of (2.6), (2.9) as

$$dP_i(t) = P_i(t)\left[r(t)\, dt + \sum_{j=1}^{d} \sigma_{ij}(t)\, d\tilde{W}_j(t) \right], \qquad i = 1, \cdots, d. \tag{2.10}$$

Comparing (2.10) with (2.2), and recalling that $\tilde{W}$ is a $\tilde{\mathbf{P}}$-Brownian motion, we can say that $\tilde{\mathbf{P}}$ is the "risk-neutral" probability measure of the market model: *it equates the appreciation rates of all the stocks to the interest rate of the bond.* Equivalently, we

[1] The only exceptions are the problems that we discuss in § 6.7 and in § 9.6.

can solve the equations (2.10), and observe that the discounted stock prices $\beta(t)P_i(t)$ with

$$\beta(t) \triangleq \frac{1}{P_0(t)} = \exp\left\{-\int_0^t r(s)\,ds\right\}, \tag{2.11}$$

are given as

$$\beta(t)P_i(t) = p_i \cdot \exp\left\{\int_0^t \sigma_i^*(s)\,d\tilde{W}(s) - \frac{1}{2}\int_0^t \|\sigma_i(s)\|^2\,ds\right\} \tag{2.12}$$

where $\sigma_i(t) = (\sigma_{i1}(t), \cdots, \sigma_{id}(t))^*$. In particular, it follows from (2.12) that *the discounted stock prices βP_i are martingales under $\tilde{\mathbf{P}}$.*

Remark 2.1. The existence of a probability measure $\tilde{\mathbf{P}}$ with the above properties will guarantee that the model is free of "arbitrage," i.e., of opportunities to make some money out of nothing (cf. Remark 4.8); on the other hand, the uniqueness of such a $\tilde{\mathbf{P}}$ will guarantee that all the risk in the market, generated by the sources of uncertainty $(W_1, \cdots, W_d)$, can be "hedged against" by skillful trading in the financial assets (Theorem 4.6 and Proposition 4.7).

These are precisely the features that will allow us to solve the option pricing (§§ 5, 6) and consumption/investment problems (§§ 8–11) in the generality of the present model.

3. A "small investor." Let us consider now an economic agent, who invests in the various securities and whose decisions cannot affect the prices in the market (a "small investor"). We shall denote by $X(t)$ the *wealth* of this agent at time t, by $\pi_i(t)$ the amount that he invests in the ith stock at that time $(1 \leqq i \leqq d)$, and by $c(t)$ the rate at which he withdraws funds for consumption.

Notice that we allow here any $\pi_i(t)$, $1 \leqq i \leqq d$ to become negative, which amounts to selling the ith stock short. Similarly, the amount of money

$$X(t) - \sum_{i=1}^{d} \pi_i(t)$$

invested in the bond at any particular time, may also become negative; this is to be interpreted as borrowing at the interest rate $r(t)$.

DEFINITION 3.1. A *portfolio process* $\pi(t) = (\pi_1(t), \cdots, \pi_d(t))^*$, $0 \leqq t \leqq T$ is an $\mathcal{R}^d$-valued process, which is progressively measurable with respect to $\{\mathcal{F}_t\}$ and satisfies

$$\int_0^T \|\pi(t)\|^2\,dt < \infty \quad \text{a.s.} \tag{3.1}$$

DEFINITION 3.2. A *consumption rate process* $c(t)$, $0 \leqq t \leqq T$ is nonnegative, progressively measurable with respect to $\{\mathcal{F}_t\}$, and satisfies

$$\int_0^T c(t)\,dt < \infty \quad \text{a.s.} \tag{3.2}$$

The adaptivity condition in Definitions 3.1 and 3.2 means of course that the investor cannot anticipate the future values of the prices; that is, "insider trading" is excluded.

With the above interpretations and notation, we obtain the following equation for the wealth $X(t)$ of the agent:

$$\begin{aligned} dX(t) &= \sum_{i=1}^{d} \pi_i(t)\left[b_i(t)\,dt + \sum_{j=1}^{d} \sigma_{ij}(t)\,dW_j(t) \right] - c(t)\,dt \\ &\quad + \left[X(t) - \sum_{i=1}^{d} \pi_i(t) \right] r(t)\,dt. \end{aligned} \tag{3.3}$$

The three terms on the right-hand side of (3.3) account, respectively, for: (a) capital gains or losses from investments in stocks, (b) the decrease in wealth due to consumption, and (c) capital gains or losses from money invested in bonds.

With the help of (2.6), (2.9) we can re-cast the wealth equation (3.3) in its vector form

$$\begin{aligned} dX(t) &= [r(t)X(t) - c(t)]\,dt + \pi^*(t)[b(t) - r(t)\mathbf{1}]\,dt + \pi^*(t)\sigma(t)\,dW(t) \\ &= [r(t)X(t) - c(t)]\,dt + \pi^*(t)\sigma(t)\,d\tilde{W}(t). \end{aligned} \tag{3.4}$$

The solution of (3.4) with initial wealth $X(0) = x \geqq 0$ is easily seen to be given, in the notation of (2.11), by

$$\beta(t)X(t) = x - \int_0^t \beta(s)c(s)\,ds + \int_0^t \beta(s)\pi^*(s)\sigma(s)\,d\tilde{W}(s), \qquad 0 \leqq t \leqq T. \tag{3.5}$$

Remark 3.3. It is easily seen from (3.5) that the process

$$M(t) = \beta(t)X(t) + \int_0^t \beta(s)c(s)\,ds, \qquad 0 \leqq t \leqq T, \tag{3.6}$$

consisting of current discounted wealth plus total discounted consumption to-date, is a continuous local martingale under $\tilde{\mathbf{P}}$. Let us introduce now the process

$$\zeta(t) = \beta(t)Z(t), \tag{3.7}$$

which modifies the discount factor of (2.11) in order to take into account the presence of the financial market. From Remark 3.3, and with the help of the "Bayes rule"

$$\tilde{\mathbf{E}}[\,Y|\mathcal{F}_s] = \frac{\mathbf{E}[\,YZ(t)|\mathcal{F}_s]}{Z(s)}, \tag{3.8}$$

which is valid for $0 \leqq s < t \leqq T$ for every $\mathcal{F}_t$-measurable, $\tilde{\mathbf{P}}$-integrable random variable Y (cf. Karatzas and Shreve (1987, p. 193)), we can deduce that the process

$$N(t) = \zeta(t)X(t) + \int_0^t \zeta(s)c(s)\,ds, \qquad 0 \leqq t \leqq T \tag{3.9}$$

is a continuous local martingale under $\mathbf{P}$, and that *N is a $\mathbf{P}$-supermartingale, if and only if M is a $\tilde{\mathbf{P}}$-supermartingale.*

Remark 3.4. The process ζ of (3.7) will play a fundamental role in the sequel. We shall see that it acts as a "deflator," in the sense that multiplication by $\zeta(t)$ converts wealth held at time t to the equivalent amount of wealth held at time zero.

It is interesting to note that (3.7), (2.11), and (2.7) lead to the linear stochastic differential equation

$$d\zeta(t) = -\zeta(t)[r(t)\,dt + \theta^*(t)\,dW(t)] \tag{3.10}$$

for the process ζ.

4. Admissible strategies. We shall single out those pairs (π, c) for which the investor avoids negative wealth at the terminal time, with probability one.

DEFINITION 4.1. A pair (π, c) of portfolio and consumption rate processes is called *admissible* for the initial capital $x \geq 0$, if the corresponding wealth process X of (3.5) satisfies

$$X(T) \geq 0 \quad \text{and} \quad \zeta(t)X(t) \geq -K, \quad \forall\, 0 \leq t \leq T \tag{4.1}$$

almost surely, for some nonnegative and $\mathbf{P}$-integrable random variable $K = K(\pi, c)$.

The class of such pairs is denoted by $\mathcal{A}(x)$.

For every $(\pi, c) \in \mathcal{A}(x)$, the continuous, $\mathbf{P}$-local martingale N of (3.9) is bounded from below; an application of Fatou's lemma shows then that N is a supermartingale under $\mathbf{P}$, and therefore that the process M of (3.6) is a supermartingale under $\tilde{\mathbf{P}}$. Consequently, with $\mathcal{S}_{u,v}$ denoting the class of $\{\mathcal{F}_t\}$-stopping times with values in the interval $[u, v]$, we have by the optional sampling theorem the equivalent inequalities

$$\mathbf{E}\left[\zeta(\tau)X(\tau) + \int_0^\tau \zeta(s)c(s)\,ds\right] \leq x, \tag{4.2}$$

$$\tilde{\mathbf{E}}\left[\beta(\tau)X(\tau) + \int_0^\tau \beta(s)c(s)\,ds\right] \leq x \tag{4.3}$$

for every $\tau \in \mathcal{S}_{0,T}$.

Remark 4.2. With the interpretation of the process ζ as a "deflator," the inequality (4.2) acquires the significance of a *budget constraint*; it mandates that "the expected total value of current wealth and consumption-to-date, both deflated down to $t = 0$, does not exceed the initial capital."

DEFINITION 4.3. For every given number $x \geq 0$, denote by

(i) $\mathcal{C}(x)$ the class of consumption rate processes c which satisfy

$$\tilde{\mathbf{E}} \int_0^T \beta(s)c(s)\,ds \leq x, \tag{4.4}$$

and by

(ii) $\mathcal{L}(x)$ the class of nonnegative, $\mathcal{F}_T$-measurable random variables B which satisfy

$$\tilde{\mathbf{E}}[\beta(T)B] \leq x. \tag{4.5}$$

From the inequality (4.3) we deduce

$$(\pi, c) \in \mathcal{A}(x) \Rightarrow c \in \mathcal{C}(x), \qquad X(T) \in \mathcal{L}(x). \tag{4.6}$$

In the next two theorems, we discuss the extent to which the opposite implications are true.

THEOREM 4.4. *For every $c \in C(x)$, there exists a portfolio process π such that $(\pi, c) \in \mathcal{A}(x)$.*

Proof. Given $c \in \mathcal{C}(x)$, introduce the random variable $D = \int_0^T \beta(s)c(s)\,ds$ and the $(\{\mathcal{F}_t\}, \tilde{\mathbf{P}})$-martingale

$$u(t) = \tilde{\mathbf{E}}(D|\mathcal{F}_t) - \tilde{\mathbf{E}}D, \qquad 0 \leq t \leq T. \tag{4.7}$$

The fundamental martingale representation theorem (e.g., Ikeda and Watanabe (1981, p. 80) or Karatzas and Shreve (1987, pp. 184, 375)) shows that u can be written as a stochastic integral with respect to $\tilde{W}$, i.e.,

$$u(t) = \int_0^t \phi^*(s)\,d\tilde{W}(s), \qquad 0 \leq t \leq T \tag{4.8}$$

for some $\{\mathcal{F}_t\}$-progressively measurable, $\mathcal{R}^d$-valued process ϕ with $\int_0^T \|\phi(s)\|^2\, ds < \infty$, almost surely. Finally, defining $\pi(t) = P_0(t)(\sigma^*(t))^{-1}\phi(t)$, we see that (4.8) is equivalently rewritten as

$$u(t) = \int_0^t \beta(s)\pi^*(s)\sigma(s)\, d\tilde{W}(s) \tag{4.9}$$

and that π is a portfolio process (i.e., (3.1) is satisfied).

The wealth process X corresponding to the pair (π, c) is given by

$$\begin{aligned} \beta(t)X(t) &= x - \int_0^t \beta(s)c(s)\, ds + u(t) \\ &= \tilde{\mathbf{E}}\left[\int_t^T \beta(s)c(s)\, ds \,\middle|\, \mathcal{F}_t\right] + x - \tilde{\mathbf{E}}D, \qquad 0 \leqq t \leqq T \end{aligned} \tag{4.10}$$

thanks to (3.5), (4.9), and (4.7). It is easily seen that this process has continuous, nonnegative paths with $X(T) = (x - \tilde{\mathbf{E}}D)P_0(T) \geqq 0$, almost surely. In other words, the pair (π, c) is admissible. □

We shall say that two measurable stochastic processes A, B on $[0, T]$ are *equivalent*, if $A(t, \omega) = B(t, \omega)$ holds for $\lambda \times \mathbf{P}$-almost everywhere $(t, \omega) \in [0, T] \times \Omega$.

Here, λ stands for Lebesgue measure.

PROPOSITION 4.5. *For every consumption rate process c in the class*

$$\mathcal{D}(x) = \left\{ c \in \mathcal{C}(x);\ \tilde{\mathbf{E}} \int_0^T \beta(s)c(s)\, ds = x \right\} \tag{4.11}$$

we have the following:

(i) *The portfolio π of Theorem* 4.4 *is unique up to equivalence.*

(ii) *The corresponding wealth process X satisfies $X(T) = 0$, almost surely.*

(iii) *The process M of* (3.6) *is a $\tilde{\mathbf{P}}$-martingale. In particular,*

$$\beta(t)X(t) = \tilde{\mathbf{E}}\left[\int_t^T \beta(s)c(s)\, ds \,\middle|\, \mathcal{F}_t\right], \qquad 0 \leqq t \leqq T.$$

Proof. For a given $c \in \mathcal{D}(x)$, and any portfolio π such that $(\pi, c) \in \mathcal{A}(x)$, we have from (4.3) the inequality

$$\tilde{\mathbf{E}}[\beta(T)X(T)] \leqq x - \tilde{\mathbf{E}} \int_0^T \beta(s)c(s)\, ds = 0,$$

which justifies (ii), as well as

$$\tilde{\mathbf{E}}M(T) = \tilde{\mathbf{E}} \int_0^T \beta(s)c(s)\, ds = x = \tilde{\mathbf{E}}M(0),$$

which establishes (iii) by showing that the supermartingale M of (3.6) has constant expectation.

Now for any two portfolios π_1, π_2 such that $(\pi_1, c) \in \mathcal{A}(x)$ and $(\pi_2, c) \in \mathcal{A}(x)$, let X_1, X_2 represent the corresponding wealth processes and M_1, M_2 the corresponding $\tilde{\mathbf{P}}$-martingales of (3.6). The martingale

$$(M_1 - M_2)(t) = \int_0^t \beta(s)(\pi_1(s) - \pi_2(s))^*\sigma(s)\, d\tilde{W}(s), \qquad 0 \leqq t \leqq T$$

is identically zero, because $M_1(T) = M_2(T) = \int_0^T \beta(s)c(s)\,ds$. Therefore,

$$\langle M_1 - M_2\rangle(t) = \int_0^t \beta^2(s)\|(\pi_1(s) - \pi_2(s))^*\sigma(s)\|^2\,ds = 0, \qquad 0 \leqq t \leqq T,$$

which shows that π_1, π_2 are equivalent. □

The following "controllability" result is analogous to Theorem 4.4, and characterizes the *levels of wealth attainable by an initial capital* $X(0) = x$.

THEOREM 4.6. *For every* $B \in \mathcal{L}(x)$, *there exists a pair* $(\pi, c) \in \mathcal{A}(x)$ *such that the corresponding wealth process* X *satisfies* $X(T) = B$, *almost surely.*

Proof. By analogy with the proof of Theorem 4.4, introduce the $\tilde{\mathbf{P}}$-martingale

$$v(t) = \tilde{\mathbf{E}}[B\beta(T)|\mathcal{F}_t] - \tilde{\mathbf{E}}[B\beta(T)], \qquad 0 \leqq t \leqq T \tag{4.12}$$

and conclude that it can be represented in the form (4.9), i.e.,

$$v(t) = \int_0^t \beta(s)\pi^*(s)\sigma(s)\,d\tilde{W}(s),$$

for a suitable portfolio π. Now the continuous process X defined by

$$\beta(t)X(t) = x - \rho t + v(t), \qquad 0 \leqq t \leqq T \tag{4.13}$$

and $\rho = (x - \tilde{\mathbf{E}}[B\beta(T)])/T$, represents the wealth corresponding to the pair (π, c) with $c(t) = \rho P_0(t)$. But then it follows from (4.12), (4.13) that

$$\beta(t)X(t) = \tilde{\mathbf{E}}[B\beta(T)|\mathcal{F}_t] + (x - \tilde{\mathbf{E}}[B\beta(T)]) \cdot \left(1 - \frac{t}{T}\right). \tag{4.14}$$

We deduce from (4.5) and (4.14) that X is nonnegative, so that in particular the pair (π, c) is admissible, and $X(T) = B$ almost surely. □

We can also establish an analogue of Proposition 4.5; we omit the easy proof.

PROPOSITION 4.7. *For any random variable* B *in the class*

$$\mathcal{M}(x) = \{B \in \mathcal{L}(x);\ \tilde{\mathbf{E}}[B\beta(T)] = x\} \tag{4.15}$$

we have the following:

(i) *The pair* (π, c) *of Theorem* 4.6 *is unique and* $c \equiv 0$, *up to equivalence.*

(ii) *The corresponding wealth process* X *is given by*

$$\beta(t)X(t) = \tilde{\mathbf{E}}[B\beta(T)|\mathcal{F}_t], \qquad 0 \leqq t \leqq T. \tag{4.16}$$

Remark 4.8. Let us define an *arbitrage opportunity* as a portfolio π such that

(i) $(\pi, 0) \in \mathcal{A}(0)$, and

(ii) The wealth process X, which corresponds to $(\pi, 0)$ and the initial capital $x = 0$, satisfies $\mathbf{P}[X(T) > 0] > 0$.

In other words, an arbitrage opportunity is an investment strategy that achieves, with zero initial capital, an amount of terminal wealth which is almost surely nonnegative and positive with positive probability. It is also sometimes called a "free lunch," for obvious reasons.

Our model excludes arbitrage opportunities; indeed, the necessary condition for admissibility (4.3) yields with $x = 0$ and $c \equiv 0$: $\tilde{\mathbf{E}}[\beta(T)X(T)] \leqq 0$, leading to $X(T) = 0$, almost surely.

Remark 4.9. Theorem 4.6 and Proposition 4.7 still hold, if T is replaced by a positive stopping time $\tau \in \mathcal{S}_{0,T}$ and B by a nonnegative, $\mathcal{F}_\tau$-measurable random variable (recall (4.3)). We would have to replace (4.14) by

$$\beta(t)X(t) = \tilde{\mathbf{E}}[B\beta(\tau)|\mathcal{F}_t] + (x - \tilde{\mathbf{E}}[B\beta(\tau)])\left(1 - \frac{t \wedge \tau}{\tau}\right)$$

and take $c(t) \equiv 0$, $\pi(t) \equiv 0$ for $\tau \leq t \leq T$. The rest of the argument goes through without change.

5. The pricing of European options. Suppose that at time $t = 0$ we sign a contract which gives us the option to buy, at the specified time T (maturity, expiration date), one share of the stock $i = 1$ at a specified price q (the contractual "exercise price"). At maturity, if the price $P_1(T)$ of the share is below the exercise price, the contract is worthless to us; on the other hand, if $P_1(T) > q$, we can exercise our option at $t = T$, buy one share of the stock at the pre-assigned price q, and then sell the share immediately in the market for $P_1(T)$.

Thus, this contract is equivalent to a payment of $(P_1(T) - q)^+$ at maturity; it is called a *European option*, in contradistinction with "American options" which can be exercised at any stopping time in $[0, T]$ (cf. § 6).

The following definition generalizes the concept of European option.

DEFINITION 5.1. A *European Contingent Claim* (ECC) is a financial instrument consisting of a payment B at maturity; here, B is a nonnegative, $\mathcal{F}_T$-measurable random variable with

$$\mathbf{E}(B^\mu) < \infty \quad \text{for some } \mu > 1. \tag{5.1}$$

Remark 5.2. Using the boundedness of the processes r and θ, as well as the Hölder inequality, it is not hard to see that (5.1) implies

$$\tilde{\mathbf{E}}[B\beta(T)] < \infty. \tag{5.2}$$

DEFINITION 5.3. A *hedging strategy* against the ECC of Definition 5.1 is a pair $(\pi, c) \in \mathcal{A}(x)$ for some $x > 0$, such that $X(T) = B$ almost surely.

We denote by $\mathcal{H}(x)$ the class of hedging strategies with initial wealth $X(0) = x$.

In words, a hedging strategy $(\pi, c) \in \mathcal{H}(x)$ starts out with initial wealth x and "reproduces the payoff from the ECC" at $t = T$.

What is a fair price to pay at $t = 0$ for the ECC? If there exists a hedging strategy for some $x > 0$, then an agent who contemplates buying the ECC at time $t = 0$ can instead invest in the market according to the portfolio π and consume at the rate c, and still achieve the same wealth at $t = T$ as the payment from the ECC. Therefore, the price he should be prepared to pay at $t = 0$ for the ECC cannot possibly be greater than this amount x.

It is natural then to define the fair price as the *smallest* value of the initial wealth, which allows the construction of a hedging strategy.

DEFINITION 5.4. The number

$$v \triangleq \inf\{x > 0;\ \exists (\pi, c) \in \mathcal{H}(x)\} \tag{5.3}$$

is called the *fair price* at $t = 0$ for the ECC of Definition 5.1.

THEOREM 5.5. *The fair price v of Definition* 5.4 *is given as*

$$v = \tilde{\mathbf{E}}\left[B \exp\left(-\int_0^T r(u)\, du \right) \right]. \tag{5.4}$$

There exists a portfolio process π with $(\pi, 0) \in \mathcal{H}(v)$; this portfolio is unique up to equivalence, and the corresponding wealth process is given by

$$X(t) = \tilde{\mathbf{E}}\left[B \exp\left(\int_t^T r(u)du \right) \middle| \mathcal{F}_t \right], \qquad 0 \leq t \leq T. \tag{5.5}$$

Proof. All the claims follow directly from Theorem 4.6 and Proposition 4.7. □

The random variable $X(t)$ of (5.5) is called the *value at time t* of the ECC.

Example 5.6. Consider a financial market model with constant interest rate $r(t) \equiv r \geqq 0$ and volatility matrix $\sigma(t) \equiv \sigma$, as well as and a contingent claim with $B = \varphi(P(T))$. Here, $\varphi : \mathcal{R}_+^d \to [0, \infty)$ is a continuous function and

$$P(t) = (P_1(t), \cdots, P_d(t))^* \tag{5.6}$$

is the vector of stock price processes which satisfy, in this case, the equations (2.10) in the form

$$dP_i(t) = P_i(t)\left[r\,dt + \sum_{j=1}^{d} \sigma_{ij}\, d\tilde{W}_j(t) \right], \qquad 1 \leqq i \leqq d. \tag{5.7}$$

The solution of these equations is given by (2.12), namely

$$P_i(t) = p_i \exp\left[\left(r - \frac{1}{2} a_{ii} \right) t + \sum_{j=1}^{d} \sigma_{ij} \tilde{W}_j(t) \right]. \tag{5.8}$$

We introduce now the function $h(t, p, y)$: $[0, \infty) \times \mathcal{R}_+^d \times \mathcal{R}^d \to \mathcal{R}_+^d$ via

$$h_i(t, p, y) \triangleq p_i \exp\left[\left(r - \frac{1}{2} a_{ii} \right) t + y_i \right], \qquad 1 \leqq i \leqq d, \tag{5.9}$$

and observe that (5.8) can be written in the vector form

$$P(t) = h(t, p, \sigma \tilde{W}(t)). \tag{5.10}$$

Coming now to the ECC with $B = \varphi(P(T))$, we see from (5.5), (5.9) that its value is given by

$$\begin{aligned} X(t) &= \tilde{\mathbf{E}}[e^{-r(T-t)} \varphi(P(T)) | \mathcal{F}_t] \\ &= \tilde{\mathbf{E}}[e^{-r(T-t)} \varphi(h(T-t, P(t), \sigma(\tilde{W}(T) - \tilde{W}(t)))) | \mathcal{F}_t] \\ &= e^{-r(T-t)} \int_{\mathcal{R}^d} \varphi(h(T-t, P(t), \sigma z)) \Gamma_{T-t}(z)\, dz \end{aligned}$$

almost surely $\tilde{\mathbf{P}}$, for every $t \in [0, T)$, where

$$\Gamma_t(z) \triangleq (2\pi t)^{-d/2} \exp\left\{ -\frac{\|z\|^2}{2t} \right\}, \qquad z \in \mathcal{R}^d, \quad t > 0$$

is the fundamental Gaussian kernel in $\mathcal{R}^d$. It follows that, with

$$v(t, p) \triangleq \begin{cases} e^{-r(T-t)} \displaystyle\int_{\mathcal{R}^d} \varphi(h(T-t, p, \sigma z)) \Gamma_{T-t}(z)\, dz, & 0 \leqq t < T, \quad p \in \mathcal{R}_+^d, \\ \varphi(p), & t = T, \quad p \in \mathcal{R}_+^d, \end{cases} \tag{5.11}$$

the value at time t of the ECC is given by

$$X(t) = v(t, P(t)). \tag{5.12}$$

In this case it is even possible to "compute" the portfolio $\pi(t)$ that achieves the value process of (5.12). Indeed, under appropriate growth conditions on φ, the function $v(t, p)$ of (5.11) is the unique solution of the Cauchy problem

$$\begin{gathered} \frac{\partial v}{\partial t} + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij} p_i p_j \frac{\partial^2 v}{\partial p_i \partial p_j} + \sum_{i=1}^{d} r p_i \frac{\partial v}{\partial p_i} - rv = 0 \quad \text{on } [0, T) \times \mathcal{R}_+^d, \\ v(T, p) = \varphi(p), \qquad p \in \mathcal{R}_+^d \end{gathered}$$

by the Feynman-Kac theorem (e.g., Karatzas and Shreve (1987, p. 366)). Applying Itô's rule to the process X of (5.12) and using the above equation and (5.7), we arrive at

$$dX(t) = rX(t)\,dt + \sum_{i=1}^{d}\sum_{j=1}^{d} \sigma_{ij}P_i(t)\frac{\partial}{\partial p_i}v(t, P(t))\,d\tilde{W}_j(t).$$

A comparison with (3.4) gives then

$$\pi_i(t) = P_i(t)\cdot\frac{\partial}{\partial p_i}v(t, P(t)), \qquad 0 \leqq t \leqq T, \quad 1 \leqq i \leqq d$$

for the portfolio process of Theorem 5.5. In other words, we should hold $N_i(t) = (\partial/\partial p_i)v(t, P(t))$ shares of the ith stock at time t.

Remark 5.7. In the particular case of a European option as in Example 5.6, with $d = 1$, $\varphi(p) = (p-q)^+$ and exercise price $q > 0$, the integration in (5.11) can be carried out in a somewhat more explicit form. Indeed, with $\Phi(z) \triangleq (1/\sqrt{2\pi})\int_{-\infty}^{z}\exp(-x^2/2)\,dx$ and

$$\nu_{\pm}(t, p; q) \triangleq \frac{1}{\sigma_{11}\sqrt{t}}\left[\log\left(\frac{p}{q}\right) + \left(r \pm \frac{1}{2}\sigma_{11}^2\right)t\right],$$

we have

(5.13)
$$v(t, p; q) = \begin{cases} p\Phi(\nu_+(T-t, p; q)) - qe^{-r(T-t)}\Phi(\nu_-(T-t, p; q)), & 0 \leqq t < T, \quad 0 < p < \infty, \\ (p-q)^+, & t = T, \quad 0 < p < \infty. \end{cases}$$

Together with

$$X(t; q) \triangleq \tilde{\mathbf{E}}[e^{-r(T-t)}(P_1(T)-q)^+ | \mathcal{F}_t] = v(t, P_1(t); q), \qquad 0 \leqq t \leqq T, \tag{5.14}$$

the expression (5.13) constitutes the celebrated *Black and Scholes* (1973) *formula.*

Note that in this formula, as well as in (5.11), the appreciation rates of the stocks do not appear; this fact makes the formulas particularly attractive and useful, because appreciation rates are usually very difficult to estimate in practice. By contrast, the interest rate $r(\cdot)$ is directly observable, and the volatilities $\sigma(\cdot)$ can in principle be estimated—albeit with some difficulty—on the basis of observations on the price processes $(P_1(t), \cdots, P_d(t))$.

More generally, any convex and piecewise C^2 function $h:[0, \infty) \to [0, \infty)$ with $h(0) = h'(0) = 0$ can be represented as

$$h(p) = \int_0^\infty (p-q)^+ h''(q)\,dq.$$

For an ECC with $B = h(P_1(T))$, the value at time t is given then by (5.5) as

$$X(t) = \tilde{\mathbf{E}}[e^{-r(T-t)}h(P_1(T))|\mathcal{F}_t] = \int_0^\infty h''(q)v(t, P_1(t); q)\,dq \qquad \text{a.s.}$$

for every $0 \leqq t \leqq T$, thanks to the Fubini theorem and (5.14).

Remark 5.8. If the expiration date T is replaced by a stopping time $\tau \in \mathcal{S}_{0,T}$ and the payment B is an $\mathcal{F}_\tau$-measurable random variable, the theory of this section still

goes through with minor changes. A hedging strategy (Definition 5.3) now has to satisfy $X(\tau) = B$ almost surely, and (5.4), (5.5) become, respectively, thanks to Remark 4.9:

$$v^{(\tau)} = \tilde{\mathbf{E}}\left[B \exp\left(-\int_0^\tau r(u)du \right) \right], \tag{5.4)'}$$

$$X^{(\tau)}(t) = \tilde{\mathbf{E}}\left[B \exp\left(-\int_t^\tau r(u)\, du \right) \middle| \mathcal{F}_t \right]. \tag{5.5)'}$$

6. The pricing of American options.[2] For the purposes of this section only, we shall need to modify slightly the model of § 3 for the small investor. First, we will have to deal with cumulative consumptions C_t up to time t, rather than with consumption rate processes.

DEFINITION 6.1. A *consumption process* $C = \{C_t;\ 0 \leqq t \leqq T\}$ is continuous, increasing, adapted to $\{\mathcal{F}_t\}$, and satisfies $C_0 = 0$, $C_T < \infty$, almost surely.

Second, we have to allow the possibility for the stocks to pay *dividends* to the stockholders, at the rate $\mu_i(t);\ 0 \leqq t \leqq T$ for every dollar invested in the ith stock, $i = 1, \cdots, d$. These are nonnegative, bounded, and $\{\mathcal{F}_t\}$-progressively measurable processes, and we denote by $\mu(t) = (\mu_1(t), \cdots, \mu_d(t))^*$ the resulting vector process.

Then the wealth process X associated to a portfolio process π (Definition 3.1) and a consumption process C (Definition 6.1) satisfies the following analogue of equations (3.3) and (3.4):

$$\begin{aligned} dX(t) &= r(t)X(t)\, dt - dC_t + \pi^*(t)[b(t) + \mu(t) - r(t)\mathbf{1}]\, dt + \pi^*(t)\sigma(t)\, dW(t) \\ &= r(t)X(t)\, dt - dC_t + \pi^*(t)\sigma(t)\, d\tilde{W}(t) \end{aligned} \tag{6.1}$$

in the notation of (2.7)-(2.9), with (2.6) replaced by

$$\theta(t) \triangleq (\sigma(t))^{-1}[b(t) + \mu(t) - r(t)\mathbf{1}]. \tag{2.6)'}$$

The notion of admissibility for a pair (π, C) remains the same as in Definition 4.1, and (4.3) becomes

$$\tilde{\mathbf{E}}\left[X(\tau)\beta(\tau) + \int_0^\tau \beta(s)\, dC_s \right] \leqq x, \qquad \forall \tau \in \mathcal{S}_{0,T} \tag{6.2}$$

for every $(\pi, C) \in \mathcal{A}(x)$.

After this setting of the stage, let us introduce the primary object of this section.

DEFINITION 6.2. An *American Contingent Claim* (ACC) is a financial instrument consisting of the following:

(i) An expiration date $T \in (0, \infty)$;

(ii) The selection of a stopping time $\tau \in \mathcal{S}_{0,T}$; and

(iii) A payoff $f(\tau)$ on exercise.

Here, $\{f(t);\ 0 \leqq t \leqq T\}$ is a continuous, nonnegative process, adapted to $\{\mathcal{F}_t\}$, which satisfies

$$\mathbf{E}\left(\sup_{0 \leqq t \leqq T} f(t) \right)^\mu < \infty \quad \text{for some } \mu > 1. \tag{6.3}$$

For instance, if $f(t) = (P_1(t) - q)^+$, we have an *American option* on the first stock that can be exercised at the price $q \geqq 0$, at any stopping time τ on $[0, T]$. We restrict attention to stopping times, in order to exclude clairvoyance.

[2] This section may be omitted on first reading, without loss of continuity; its results will not be used in the sequel.

As in § 5, we are interested in the following *pricing problem* for the ACC: What is a fair price to pay at $t=0$ for this instrument? How much is it worth at any later time $t \in (0, T]$?

Let us suppose for a moment that the selection of $\tau \in \mathcal{S}_{0,T}$ ((ii) in Definition 6.2) has been made; then we have, from Remark 5.8, the expressions

$$X^{(\tau)}(t) = \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_t^\tau r(s)\, ds \right) \middle| \mathcal{F}_t \right],$$

$$v^{(\tau)} = X^{(\tau)}(0) = \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_0^\tau r(s)\, ds \right) \right]$$

for the value of the claim and for its fair price at $t=0$. It is conceivable then that, in order to find the corresponding quantities for the ACC, we would merely have to maximize over stopping times. In particular, we should expect the fair price at $t=0$ to be given by

$$\sup_{\tau \in S_{0,T}} \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_0^\tau r(s)\, ds \right) \right],$$

and the value of the ACC at any time $t \in [0, T]$ by

$$\operatorname*{ess\,sup}_{\tau \in \mathcal{S}_{t,T}} \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_t^\tau r(s)\, ds \right) \middle| \mathcal{F}_t \right] \quad \text{a.s.}$$

The question is whether this process is the wealth corresponding to an admissible portfolio/consumption process pair, that somehow again *duplicates* the payoff from the contingent claim and does so with minimal initial capital.

DEFINITION 6.3. Fix $x > 0$; a pair $(\pi, C) \in \mathcal{A}(x)$ is called a *hedging strategy* against the ACC with initial wealth x, if the corresponding wealth process X satisfies

(i) $X(t) \geqq f(t)$, for all $0 \leqq t \leqq T$,

(ii) $X(T) = f(T)$,

almost surely. We denote by $\hat{\mathcal{H}}(x)$ the collection of all such pairs.

DEFINITION 6.4. The number

$$\hat{v} \triangleq \inf\{x > 0;\ \exists (\pi, C) \in \hat{\mathcal{H}}(x)\} \tag{6.4}$$

is called the *fair price* for the ACC of Definition 6.2.

For every $(\pi, C) \in \hat{\mathcal{H}}(x)$, we have from (6.2): $\tilde{\mathbf{E}}[f(\tau)\beta(\tau)] \leqq x$, for all $\tau \in \mathcal{S}_{0,T}$. Therefore, with

$$u(t) \triangleq \sup_{\tau \in \mathcal{S}_{t,T}} \tilde{\mathbf{E}} Q(\tau), \quad Q(t) = f(t)\beta(t), \quad 0 \leqq t \leqq T, \tag{6.5}$$

we obtain $u(0) \leqq x$, whence

$$u(0) \leqq \hat{v}. \tag{6.6}$$

We shall show that equality actually holds in (6.6).

THEOREM 6.5. *The fair price $\hat{v}$ of Definition 6.4 is given by*

$$\hat{v} = u(0) = \sup_{\tau \in \mathcal{S}_{0,T}} \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_0^\tau r(s)\, ds \right) \right]. \tag{6.7}$$

There exists a pair $(\hat{\pi}, \hat{C}) \in \hat{\mathcal{H}}(u(0))$, such that the corresponding wealth process $\hat{X}$ is given as

$$\hat{X}(t) = \operatorname*{ess\,sup}_{\tau \in \mathcal{S}_{t,T}} \tilde{\mathbf{E}}\left[f(\tau) \exp\left(-\int_t^\tau r(s)\, ds \right) \middle| \mathcal{F}_t \right] \quad \text{a.s.} \tag{6.8}$$

for every $t \in [0, T]$, *and*

$$\int_0^T 1_{\{\hat{X}(t) > f(t)\}}\, d\hat{C}_t = 0 \quad \text{a.s.} \tag{6.9}$$

holds.

In view of (6.6), only the second claim needs to be discussed. For this purpose, we have to recall some basic facts from the *theory of optimal stopping* for a continuous process such as Q (cf. Fakeev (1970), Bismut and Skalli (1977), El Karoui (1981)). We know from these sources that there exists a nonnegative, RCLL (Right-Continuous with Left-hand Limits) $\tilde{\mathbf{P}}$-supermartingale $\{Y(t), \mathcal{F}_t;\ 0 \leqq t \leqq T\}$, such that the function $u(\cdot)$ of (6.5) is given as

$$u(t) = \tilde{\mathbf{E}}\, Y(t), \qquad 0 \leqq t \leqq T, \tag{6.10}$$

and

$$Y(t) = \operatorname*{ess\,sup}_{\tau \in \mathcal{S}_{t,T}} \tilde{\mathbf{E}}[Q(\tau) | \mathcal{F}_t] \quad \text{a.s.} \tag{6.11}$$

holds for every $t \in [0, T]$. Y is the *Snell envelope* of Q, i.e., the smallest RCLL supermartingale that dominates Q, and provides the optimal stopping time τ_t for the problem of (6.5): $u(t) = \tilde{\mathbf{E}} Q(\tau_t)$, in the form

$$\tau_t \triangleq \inf\{s \in [t, T];\ Y(s) = Q(s)\}. \tag{6.12}$$

Using (6.3) and the Doob and Jensen inequalities, it can be shown that Y *is of class* $D[0, T]$ *under* $\tilde{\mathbf{P}}$, i.e., that

$$\{Y(\tau)\}_{\tau \in \mathcal{S}_{0,T}} \quad \text{is a } \tilde{\mathbf{P}}-\text{uniformly integrable family.} \tag{6.13}$$

Bismut and Skalli (1977) also show that Y *is regular*:

$$\begin{aligned} &\text{For every monotone sequence } \{\sigma_n\}_{n=1}^{\infty} \subseteq \mathcal{S}_{0,T} \text{ with} \\ &\lim_{n\to\infty} \sigma_n = \sigma \in \mathcal{S}_{0,T}, \text{ we have } \lim_{n\to\infty} \tilde{\mathbf{E}}\, Y(\sigma_n) = \tilde{\mathbf{E}}\, Y(\sigma). \end{aligned} \tag{6.14}$$

Proof of Theorem 6.5. From (6.13), (6.14) we conclude that Y admits the Doob-Meyer decomposition (e.g., Karatzas and Shreve (1987, § 1.4)):

$$Y(t) = u(0) + M(t) - A(t), \qquad 0 \leqq t \leqq T,$$

where $\{M(t), \mathcal{F}_t\}$ is a $\tilde{\mathbf{P}}$-martingale and A is a *continuous*, nondecreasing process, with $M(0) = A(0) = 0$. As in the proof of Theorem 4.4, we have the representation

$$M(t) = \int_0^t \beta(s) \hat{\pi}^*(s) \sigma(s)\, d\tilde{W}(s), \qquad 0 \leqq t \leqq T$$

of the martingale M as a stochastic integral with respect to $\tilde{W}$, for a suitable portfolio process $\hat{\pi}$. Now define

$$\hat{X}(t) = Y(t) P_0(t), \qquad 0 \leqq t \leqq T, \tag{6.15}$$

and apply Itô's rule to obtain

$$d\hat{X}(t) = r(t) \hat{X}(t)\, dt - d\hat{C}_t + \hat{\pi}^*(t) \sigma(t)\, d\tilde{W}(t)$$

for the choice

$$\hat{C}_t = \int_0^t P_0(s)\, dA(s). \tag{6.16}$$

In other words, $\hat{X}$ is the wealth process corresponding to the portfolio/consumption process pair $(\hat{\pi}, \hat{C})$, which is easily seen to belong to $\hat{\mathcal{H}}(u(0))$. The representation (6.8) follows from (6.11), and (6.9) from

$$\int_0^T 1_{\{Y(t)>Q(t)\}}\, dA(t) = 0 \quad \text{a.s.}$$

(cf. Bismut and Skalli (1977), El Karoui (1981)). □

The stopping time τ_t of (6.12) can be written equivalently as

$$\tau_t = \inf\{s \in [t, T]; \quad \hat{X}(s) = f(s)\} \quad \text{a.s.}; \tag{6.17}$$

obviously, τ_0 *provides the optimal exercise time for the* ACC. The random variable $\hat{X}(t)$ of (6.8) gives the value of the ACC at time t.

Remark 6.6. Suppose that the process Q of (6.5) is a submartingale under $\tilde{P}$; then $u(t) = \tilde{\mathbf{E}}Q(T)$, $\tau_t = T$ for every $0 \leqq t \leqq T$, and *the* ACC *should not be exercised before the expiration date* (i.e., is equivalent to an ECC).

For instance, in the setting of Example 5.6, suppose that the function $\varphi : \mathcal{R}_+^d \to [0, \infty)$ is of class C^2 and satisfies

$$\frac{1}{2}\sum_{i=1}^d \sum_{j=1}^d a_{ij} p_i p_j \frac{\partial^2 \varphi(p)}{\partial p_i \partial p_j} + \sum_{i=1}^d (r - \mu_i) p_i \frac{\partial \varphi(p)}{\partial p_i} \geqq r\varphi(p)$$

as well as a polynomial growth condition. Then $Q(t) = e^{-rt}\varphi(P(t))$ is a $\tilde{\mathbf{P}}$-submartingale, and the value process for the ACC with $f(t) = \varphi(P(t))$ is given by (5.12), with the understanding that r has to be replaced by $r - \mu_i$ in the expressions (5.7)-(5.9).

As another example, take the American option with $d = 1$, $f(t) = (P_1(t) - q)^+$, $q > 0$ written on a stock which pays no dividends: $\mu_1(t) \equiv 0$, $r(t) \geqq 0$. Then

$$Q(t) = (P_1(t)\beta(t) - q\beta(t))^+$$

is a $\tilde{\mathbf{P}}$-submartingale, and we recover a result of Merton (1973): *an American option with positive exercise price, written on a stock that pays no dividends, should not be exercised before the expiration date.*

Remark 6.7. *The infinite horizon case.* In the setting of Example 5.6 and with $\mu_i(t) \equiv \mu$, $f(t) = \varphi(P(t))$, the value process $\hat{X}$ of (6.8) for $T = \infty$ is given, formally at least, as

$$\hat{X}(t) = v(P(t)), \qquad 0 \leqq t < \infty, \tag{6.18}$$

where $v : \mathcal{R}_+^d \to [0, \infty)$ is the least r-excessive majorant of the function φ (Fakeev (1971)).

More specifically, if $d = 1$ and $\varphi(p) = (p-1)^+$, the function v of (6.18) was computed by McKean (1965) as

$$v(p) = \begin{cases} (\kappa - 1)(p/\kappa)^\gamma; & 0 < p < \kappa \\ p - 1; & \kappa \leqq p < \infty \end{cases}$$

with $\gamma = (1/\sigma^2)(\sqrt{\delta^2 + 2r\sigma^2} - \delta)$, $\alpha = r - \mu > 0$, $\delta = \alpha - \sigma^2/2$, $\kappa = \gamma/(\gamma - 1) > 1$, and the optimal exercise time of (6.17) is given by

$$\tau_t = \inf\{s \geqq t;\ P_1(s) \geqq \kappa\}.$$

The finite-horizon version of this problem is studied in Van Moerbeke (1976); we then face a genuine *free-boundary problem*, for a moving boundary $\kappa(t)$, $0 \leqq t \leqq T$ rather than a point κ as above. Van Moerbeke studies this question by reducing it to a free-boundary problem of the Stefan type.

7. Utility functions. To formulate meaningful optimization problems for the small investor of § 3, we shall need the concept of utility function.

Let $U:[0, T]\times(0, \infty)\to\mathcal{R}$ be a $C^{0,1}$ function with the following properties for each $t\in[0, T]$:

(i) $U(t, \cdot)$ is strictly increasing and strictly concave;

(ii) The derivative $U'(t, c)\triangleq(\partial/\partial c)U(t, c)$ satisfies $\lim_{c\to\infty}U'(t, c)=0$ and $U'(t, 0+)\triangleq\lim_{c\downarrow 0}U'(t, c)=\infty$.

A function with these properties will be called a *utility function.*

Remark 7.1. The assumption (i) says that the investor prefers higher levels of consumption and/or terminal wealth to lower levels (strict increase of $U(t, \cdot)$), but that he is also *risk-averse*, i.e., that his marginal utility $U'(t, c)$ is decreasing in the argument c (strict concavity of $U(t, \cdot)$) and tends to zero as $c\to\infty$ (a "saturation effect").

The assumption $U'(t, 0+)=\infty$ of (ii) is not necessary, and is imposed here only for simplicity of exposition; in the optimization problems of §§ 8, 9 it will guarantee that the constraint $c(t)\geqq 0$ on consumption (respectively, $X(T)\geqq 0$ on terminal wealth) will never be active.

We shall denote by $I(t, \cdot)$ the inverse of the strictly decreasing mapping $U'(t, \cdot)$ from $(0, \infty)$ onto itself. The inequality

$$U(t, I(t, y))\geqq U(t, c)+y[I(t, y)-c], \quad \forall c\geqq 0, \tag{7.1}$$

valid for every $(t, y)\in[0, T]\times(0, \infty)$, is then an elementary consequence of the concavity of $U(t, \cdot)$.

For certain of our results we shall need to impose the additional conditions

$$U(t, \cdot)\in C^2((0, \infty)), \quad \forall t\in[0, T], \tag{7.2}$$

$$U''(t, c)\triangleq\frac{\partial^2}{\partial c^2}U(t, c) \quad \text{is nondecreasing in } c\in(0, \infty) \text{ for all } t\in[0, T]. \tag{7.3}$$

Under (7.2) and (7.3), $I(t, \cdot)$ is convex and of class C^1 on $(0, \infty)$, and we have

$$\frac{\partial}{\partial y}U(t, I(t, y))=y\frac{\partial}{\partial y}I(t, y), \quad \forall y\in(0, \infty). \tag{7.4}$$

8. Maximization of utility from consumption. In this section we shall try to address the following question. How should a small investor, endowed with initial wealth $x>0$, choose at every time his stock portfolio $\pi(t)$ and his consumption rate $c(t)$, from among admissible pairs $(\pi, c)\in\mathcal{A}(x)$, in order to obtain a maximum expected utility from consumption?

In order to give a precise meaning to this question, let us consider a utility function U_1, and try to maximize the *expected utility from consumption*

$$J_1(x; \pi, c)=\mathbf{E}\int_0^T U_1(t, c(t))\,dt \tag{8.1}$$

over the class $\mathcal{A}_1(x)$ of pairs $(\pi, c)\in\mathcal{A}(x)$ which satisfy

$$\mathbf{E}\int_0^T U_1^-(t, c(t))\,dt<\infty. \tag{8.2}$$

We shall denote by

$$V_1(x)=\sup_{(\pi, c)\in\mathcal{A}_1(x)}J_1(x; \pi, c) \tag{8.3}$$

the *value function* of this problem.

Now according to Theorem 4.4, *the problem* (8.3) *amounts to maximizing the expression* $\mathbf{E}\int_0^T U_1(t, c(t))\,dt$ *subject to* (8.2) *and the requirement* $c \in \mathscr{C}(x)$:

$$\tilde{\mathbf{E}}\int_0^T \beta(t)c(t)\,dt = \mathbf{E}\int_0^T \zeta(t)c(t)\,dt \leqq x, \tag{8.4}$$

where ζ is the process of (3.7).

But this question is straightforward, and concerns only the consumption process; elementary Lagrange multiplier considerations suggest that the optimal c should satisfy $U_1'(t, c(t)) = y\zeta(t)$, or equivalently

$$c(t) = I_1(t, y\zeta(t)), \qquad 0 \leqq t \leqq T \tag{8.5}$$

for an appropriate constant $y > 0$. This latter should be determined so that the requirement (8.4) is satisfied as an *equality*, i.e.,

$$\mathbf{E}\int_0^T \zeta(t)I_1(t, y\zeta(t))\,dt = x, \tag{8.6}$$

because we are trying to maximize total expected utility from consumption, and this utility increases as the consumption increases.

Let us now substantiate the heuristics of the preceding paragraph. We start by introducing the function

$$\mathscr{X}_1(y) = \mathbf{E}\int_0^T \zeta(t)I_1(t, y\zeta(t))\,dt, \qquad 0 < y < \infty, \tag{8.7}$$

and assuming that

$$\mathscr{X}_1(y) < \infty, \quad \forall y \in (0, \infty). \tag{8.8}$$

It is not hard to show that $\mathscr{X}_1 : (0, \infty) \to (0, \infty)$ *is continuous and strictly decreasing, with* $\mathscr{X}_1(0+) = \infty$ *and* $\mathscr{X}_1(\infty) = 0$. Therefore, $\mathscr{X}_1$ has the inverse $\mathscr{Y}_1 = \mathscr{X}_1^{-1}$, and there is exactly one number $y = \mathscr{Y}_1(x_1)$ that satisfies (8.6) for any given $x = x_1 > 0$. We consider then the corresponding consumption process in (8.5), namely

$$c_1(t) = I_1(t, \mathscr{Y}_1(x_1) \cdot \zeta(t)), \qquad 0 \leqq t \leqq T. \tag{8.9}$$

By construction, c_1 belongs to the class $\mathscr{D}(x_1)$ of (4.11), and according to Proposition 4.5 there exists a unique (up to equivalence) portfolio process π_1 such that $(\pi_1, c_1) \in \mathscr{A}(x_1)$; the wealth process X_1 corresponding to this pair is given by

$$\begin{aligned} \beta(t)X_1(t) &= \tilde{\mathbf{E}}\left[\int_t^T \beta(s)c_1(s)\,ds \,\middle|\, \mathscr{F}_t\right] \\ &= x_1 - \int_0^t \beta(s)c_1(s)\,ds + \int_0^t \beta(s)\pi_1^*(s)\sigma(s)\,d\tilde{W}(s). \end{aligned} \tag{8.10}$$

In particular, X_1 is positive on $[0, T)$ and vanishes at $t = T$, almost surely.

THEOREM 8.1. *Assume that* (8.8) *holds. Then for any* $x_1 > 0$ *and with* c_1 *given by* (8.9), *the above pair* (π_1, c_1) *belongs to* $\mathscr{A}_1(x_1)$ *and is optimal for the problem of* (8.3):

$$V_1(x_1) = \mathbf{E}\int_0^T U_1(t, c_1(t))\,dt. \tag{8.11}$$

Proof. We need to show that c_1 satisfies (8.2), and that for any other $c \in \mathscr{C}(x_1)$ which satisfies this condition we have the comparison

$$\mathbf{E}\int_0^T U_1(t, c_1(t))\,dt \geqq \mathbf{E}\int_0^T U_1(t, c(t))\,dt. \tag{8.12}$$

Now for any such c, the inequality (7.1) gives

$$U_1(t, c_1(t)) \geqq U_1(t, c(t)) + \mathscr{Y}_1(x_1)[\zeta(t) I_1(t, \mathscr{Y}_1(x_1)\zeta(t)) - \zeta(t)c(t)], \qquad 0 \leqq t \leqq T. \tag{8.13}$$

The constant consumption $c(t) \equiv \hat{c} \triangleq x_1/\mathbf{E}\int_0^T \zeta(s)\,ds$ belongs to $\mathscr{D}(x_1)$, and for this choice the right-hand side of (8.13) is $\lambda \times \mathbf{P}$-integrable (the value of its integral is actually $\int_0^T U_1(t, \hat{c})\,dt$). It follows that c_1 satisfies (8.2).

Now for any $c \in \mathscr{C}(x_1)$ satisfying (8.2), integrate both members of (8.13) with respect to $\lambda \times \mathbf{P}$, to obtain

$$\mathbf{E}\int_0^T U_1(t, c_1(t))\,dt \geqq \mathbf{E}\int_0^T U_1(t, c(t))\,dt + \mathscr{Y}_1(x_1)\cdot\left[x_1 - \mathbf{E}\int_0^T \zeta(t)c(t)\,dt\right].$$

The expression in the brackets is nonnegative, and (8.12) follows. □

In order to characterize the value function V_1 of (8.3) a little more precisely, we study the expected utility associated with a consumption rate process of the form (8.5), namely

$$G_1(y) \triangleq \mathbf{E}\int_0^T U_1(t, I_1(t, y\zeta(t)))\,dt, \qquad y \in (0, \infty) \tag{8.14}$$

under the assumption

$$\mathbf{E}\int_0^T |U_1(t, I_1(t, y\zeta(t)))|\,dt < \infty, \quad \forall y \in (0, \infty). \tag{8.15}$$

Then G_1 is a continuous, strictly decreasing function, and from Theorem 8.1 the value function of (8.3) is obviously given as

$$V_1 = G_1 \circ \mathscr{Y}_1. \tag{8.16}$$

Furthermore, formal differentiations in (8.7), (8.14) yield

$$\mathscr{X}_1'(y) = \mathbf{E}\int_0^T \zeta^2(t)\frac{\partial}{\partial y} I_1(t, y\zeta(t))\,dt$$

and

$$G_1'(y) = E\int_0^T \zeta(t) U_1'(I_1(t, y\zeta(t)))\frac{\partial}{\partial y} I_1(t, y\zeta(t))\,dt = y\mathscr{X}_1'(y).$$

These formalities can be made rigorous under the conditions (7.2), (7.3) and their consequence (7.4). We then arrive at the following characterization.

PROPOSITION 8.2. *Under the conditions* (8.8), (8.15) *on the utility function* U_1, *the value* $V_1(\cdot)$ *of the problem* (8.3) *is given by* (8.16).

Furthermore, if (7.2) *and* (7.3) *are also satisfied by* U_1, *then the strictly decreasing functions* $\mathscr{X}_1$ *and* G_1 *of* (8.7), (8.14) *are also continuously differentiable, and we have* $G_1'(y) = y\mathscr{X}_1'(y)$, *for all* $0 < y < \infty$ *as well as its consequence*

$$V_1' = \mathscr{Y}_1 \tag{8.17}$$

from (8.16); *in particular,* V_1 *is strictly increasing and strictly concave.*

Example 8.3. In the important special case $U_1(t, c) = \exp\{-\int_0^t \mu(s)\, ds\} \cdot \log c$ with bounded, measurable $\mu : [0, T] \to \mathcal{R}$, we obtain

$$\mathscr{X}_1(y) = \frac{\alpha_1}{y}, \qquad G_1(y) = -\alpha_1 \cdot \log y + \delta_1 \tag{8.18}$$

and hence

$$V_1(x) = \alpha_1 \cdot \log\left(\frac{x}{\alpha_1}\right) + \delta_1,$$

where

$$\begin{aligned} \alpha_1 &= \int_0^T \exp\left(-\int_0^t \mu(s)\, ds\right) dt, \\ \delta_1 &= \mathbf{E}\int_0^T \exp\left(-\int_0^t \mu(s)\, ds\right)\left\{\int_0^t \left(r(s) + \frac{1}{2}\|\theta(s)\|^2 - \mu(s)\right) ds\right\} dt. \end{aligned} \tag{8.19}$$

In particular, (8.8) and (8.15) are satisfied trivially in this case.

Example 8.4. In the special case $U_1(t, c) = -\exp\{-\int_0^t \mu(s)\, ds\}/c$, with μ as in Example 8.3, we obtain

$$\mathscr{X}_1(y) = \alpha_1 y^{-1/2}, \qquad G_1(y) = -\alpha_1 y^{1/2}, \tag{8.20}$$

and thus $V_1(x) = -\alpha_1^2/x$, where now

$$\alpha_1 = \mathbf{E}\int_0^T \exp\left(-\frac{1}{2}\int_0^t (\mu(s) + r(s))\, ds\right) Z^{1/2}(t)\, dt. \tag{8.21}$$

Again, (8.8) and (8.15) are obviously satisfied in this case.

Remark 8.5. If $U_1(0) > -\infty$, we have $\mathscr{A}_1(x) = \mathscr{A}(x)$, and it can be shown easily that (8.15) implies (8.8).

9. Maximization of utility from investment. Let us take up now the complementary problem to that of § 8, namely the maximization of the *expected utility from terminal wealth*

$$J_2(x; \pi, c) = \mathbf{E}\, U_2(T, X(T)), \tag{9.1}$$

over the class $\mathscr{A}_2(x)$ of pairs $(\pi, c) \in \mathscr{A}(x)$ that satisfy

$$\mathbf{E}\, U_2^-(T, X(T)) < \infty. \tag{9.2}$$

Here, U_2 is a utility function as in § 7, and

$$V_2(x) \triangleq \sup_{(\pi, c) \in \mathscr{A}_2(x)} J_2(x; \pi, c) \tag{9.3}$$

is the *value function* of this problem.

In this setting the agent obviously tries to maximize the utility from his terminal wealth, within the constraints imposed by the level of his initial capital and quantified by the *budget constraint* (4.5), i.e.,

$$\tilde{\mathbf{E}}[\beta(T) X(T)] = \mathbf{E}[\zeta(T) X(T)] \leqq x, \tag{9.4}$$

which mandates that "the expected terminal wealth, deflated down to $t = 0$, should not exceed the initial capital."

According to Theorem 4.6, *the problem* (9.3) *amounts to maximizing the expression* $\mathbf{E}\, U_2(T, X(T))$ *over the class of nonnegative,* $\mathscr{F}_T$*-measurable random variables* $X(T)$ *that satisfy* (9.2) *and* (9.4).

The situation is completely analogous to that of the previous section, so we just outline the results. We introduce the decreasing function

$$\mathscr{X}_2(y) = \mathbf{E}[\zeta(T) I_2(T, y\zeta(T))], \qquad 0 < y < \infty, \tag{9.5}$$

assume that

$$\mathscr{X}_2(y) < \infty, \qquad \forall y \in (0, \infty), \tag{9.6}$$

and show that $\mathscr{X}_2 : (0, \infty) \to (0, \infty)$ is continuous and strictly decreasing with $\mathscr{X}_2(0+) = \infty$, $\mathscr{X}(\infty) = 0$. Denoting by $\mathscr{Y}_2 = \mathscr{X}_2^{-1}$ the inverse of this function, and fixing an initial capital $x = x_2 > 0$, we introduce the $\mathscr{F}_T$-measurable random variable

$$X_2(T) = I_2(T, \mathscr{Y}_2(x_2)\zeta(T)) \tag{9.7}$$

and observe that it belongs to the class $\mathscr{M}(x_2)$ of (4.15). From Proposition 4.7, there exists a unique (up to equivalence) pair $(\pi_2, c_2) \in \mathscr{A}(x_2)$ that almost surely achieves the terminal wealth of (9.7); for this pair we have $c_2 \equiv 0$, and the corresponding wealth process X_2 is given by

$$\begin{aligned} \beta(t) X_2(t) &= \tilde{\mathbf{E}}[\beta(T) X_2(T) | \mathscr{F}_t] \\ &= x_2 + \int_0^t \beta(s) \pi_2^*(s) \sigma(s)\, d\tilde{W}(s), \qquad 0 \leqq t \leqq T. \end{aligned} \tag{9.8}$$

THEOREM 9.1. *Under the assumption* (9.6), *fix* $x_2 > 0$ *and consider the random variable* $X_2(T)$ *of* (9.7). *Then the above pair* $(\pi_2, 0)$ *belongs to* $\mathscr{A}_2(x_2)$ *and achieves the supremum in* (9.3):

$$V_2(x_2) = \mathbf{E} U_2(T, X_2(T)). \tag{9.9}$$

Sketch of Proof. Using the inequality (7.1) we show that the random variable $X_2(T)$ of (9.7) satisfies (9.2), and that

$$\mathbf{E} U_2(T, X_2(T)) \geqq \mathbf{E} U_2(T, X(T)) \tag{9.10}$$

holds for any other random variable $X(T) \in \mathscr{L}(x_2)$ satisfying (9.2). The details are completely analogous to those in the Proof of Theorem 8.1, and are left to the reader. □

We also have the following characterization of the value function.

PROPOSITION 9.2. *Under the conditions* (9.6) *and*

$$\mathbf{E}|U_2(T, I_2(T, y\zeta(T))| < \infty, \quad \forall y \in (0, \infty) \tag{9.11}$$

on the utility function U_2, *the value* V_2 *of the problem* (9.3) *is given as*

$$V_2 = G_2 \circ \mathscr{Y}_2, \tag{9.12}$$

where G_2 *is the continuous, strictly decreasing function*

$$G_2(y) \triangleq \mathbf{E} U_2(T, I_2(T, y\zeta(T))), \qquad 0 < y < \infty. \tag{9.13}$$

Furthermore, if (7.2) *and* (7.3) *are also satisfied by* U_2, *then the functions* $\mathscr{X}_2$, G_2 *of* (9.5), (9.13) *are also continuously differentiable, and satisfy* $G_2'(y) = y\mathscr{X}_2'(y)$ *for all* $0 < y < \infty$. *In that case we have*

$$V_2' = \mathscr{Y}_2, \tag{9.14}$$

which implies that V_2 *is strictly increasing and strictly concave.*

Example 9.3. In the special case $U_2(T, c) = \exp\{-\int_0^T \mu(s)\, ds\} \log c$, with μ as in Example 8.3, we deduce easily

$$\mathscr{X}_2(y) = \frac{\alpha_2}{y}, \quad G_2(y) = -\alpha_2 \cdot \log y + \delta_2, \quad \text{and}$$
$$V_2(x) = \alpha_2 \cdot \log\left(\frac{x}{\alpha_2}\right) + \delta_2, \tag{9.15}$$

where

$$\alpha_2 = \exp\left(\int_0^T \mu(s)\, ds\right),$$
$$\delta_2 = \mathbf{E}\left[\exp\left(-\int_0^T \mu(s)\, ds\right)\left\{\int_0^T \left(r(s) + \frac{1}{2}\|\theta(s)\|^2 - \mu(s)\right) ds\right\}\right]. \tag{9.16}$$

In particular, take $\mu \equiv 0$; then $I_2(T, y) = \mathscr{X}_2(y) = 1/y$, and from (9.7) we obtain

$$\beta(T) X_2(T) = x_2 \cdot \exp\left\{\int_0^T \theta^*(s)\, dW(s) + \frac{1}{2}\int_0^T \|\theta(s)\|^2\, ds\right\}.$$

On the other hand, (9.8) gives the optimal wealth process X_2 as

$$\beta(t) X_2(t) = x_2 \cdot \exp\left\{\int_0^t \theta^*(s)\, dW(s) + \frac{1}{2}\int_0^t \|\theta(s)\|^2\, ds\right\}, \tag{9.17}$$

and it follows from an easy application of Itô's rule that the process βX_2 satisfies the linear wealth equation

$$\beta(t) X_2(t) = x_2 + \int_0^t \beta(s) X_2(s) \theta^*(s)[dW(s) + \theta(s)\, ds]. \tag{9.18}$$

A comparison with (9.8) shows that *the optimal portfolio for maximizing* $E[\log X(T)]$ *is given explicitly as*

$$\pi_2(t) = X_2(t)(\sigma^*(t))^{-1}\theta(t). \tag{9.19}$$

Note that the processes X_2, π_2 can be defined by (9.17), (9.19) on the entirety of $[0, \infty)$, and that the above maximization holds then for every finite $T > 0$.

Remark 9.4. A similar analysis can be carried out in the context of Example 8.3 with $\mu \equiv 0$; it leads to the explicit computations $X_1(t) = x_1(T-t)/T\zeta(t)$, $c_1(t) = X_1(t)/(T-t)$, and $\pi_1(t) = X_1(t)(\sigma^*(t))^{-1}\theta(t)$ for the optimal wealth, consumption, and portfolio processes. We leave the details to the care of the diligent reader.

Example 9.5. In the special case $U_2(T, c) = -\exp\{-\int_0^T \mu(s)\, ds\}/c$, with μ as in Example 8.3, we have

$$\mathscr{X}_2(y) = \alpha_2 y^{-1/2}, \quad G_2(y) = -\alpha_2 y^{1/2}, \quad V_2(x) = -\frac{\alpha_2^2}{x}, \tag{9.20}$$

where

$$\alpha_2 = \mathbf{E}\left[\exp\left\{-\frac{1}{2}\int_0^T (\mu(s) + r(s))\, ds\right\} \cdot Z^{1/2}(T)\right]. \tag{9.21}$$

9.6. Maximizing the growth rate from investment. For the purposes of this paragraph only, let us call an $\{\mathcal{F}_t\}$-progressively measurable process $\pi(t,\omega):[0,\infty)\times\Omega\to\mathcal{R}^d$ an *admissible portfolio*, if it satisfies (3.1) for every finite $T>0$, and if the wealth process X corresponding to π and zero consumption, i.e.,

$$\beta(t)X(t)=x_2+\int_0^t \beta(s)\pi^*(s)\sigma(s)[dW(s)+\theta(s)\,ds], \qquad 0\leqq t<\infty, \tag{9.22}$$

is nonnegative, almost surely. Certainly the portfolio π_2 of (9.19) is admissible, since the corresponding wealth process X_2 in (9.18) is actually positive.

Quite obviously, π_2 *maximizes the expected growth rate* $\underline{\lim}_{T\to\infty}(1/T)\,E[\log X(T)]$ *from investment*; indeed, we noticed in Example 9.3 that $E[\log X(T)]\leqq E[\log X_2(T)]$ holds for every finite $T>0$, where X is the wealth process corresponding to an arbitrary admissible portfolio π. It then follows that

$$\begin{aligned}\varliminf_{T\to\infty}\frac{1}{T}E[\log X(T)]&\leqq\varliminf_{t\to\infty}\frac{1}{T}E[\log X_2(T)]\\&=\varliminf_{T\to\infty}\frac{1}{T}\int_0^T E\left\{r(s)+\frac{1}{2}\|\theta(s)\|^2\right\}ds.\end{aligned} \tag{9.23}$$

Consider now the problem of *maximizing the actual growth rate* $\underline{\lim}_{T\to\infty}(1/T)\log X(T)$, *over admissible portfolios* π. The comparison (9.23) suggests π_2 as a very strong candidate for this problem as well. In fact, we shall show that the comparison

$$\varlimsup_{T\to\infty}\frac{1}{T}\log X(T)\leqq\varliminf_{T\to\infty}\frac{1}{T}\log X_2(T)=\varliminf_{T\to\infty}\frac{1}{T}\int_0^T\left\{r(s)+\frac{1}{2}\|\theta(s)\|^2\right\}ds \tag{9.24}$$

holds almost surely, for every admissible portfolio π and its associated wealth process X.

The equality in (9.24) follows easily from (9.17) and the fact that $\lim_{T\to\infty}(1/T)\int_0^T\theta^*(s)\,dW(s)=0$, almost surely. To obtain the inequality, we apply Itô's rule to the ratio $\Lambda(t)=X(t)/X_2(t)$; in conjunction with (9.22) and (9.18), this leads to

$$d\Lambda(t)=X_2^{-1}(t)[\pi^*(t)\sigma(t)-X(t)\theta^*(t)]\,dW(t)$$

and shows that Λ is a nonnegative local martingale, hence a supermartingale. As a nonnegative supermartingale, Λ has a last element $\Lambda(\infty)\triangleq\lim_{t\to\infty}\Lambda(t)$, and satisfies the inequality

$$e^{\delta n}P[\sup_{n\leqq t<\infty}\Lambda(t)>e^{\delta n}]\leqq E\Lambda(n)\leqq 1$$

for every integer $n\geqq 1$ and $\delta>0$ (cf. Karatzas and Shreve (1987, Problem 1.3.16, Theorem 1.3.8)). It follows that

$$\sum_{n=1}^{\infty}P[\sup_{n\leqq t<\infty}\log\Lambda(t)>\delta n]\leqq\sum_{n=1}^{\infty}e^{-\delta n}<\infty,$$

and by the Borel-Cantelli Lemma there exists an integer-valued random variable N_δ such that, for almost every $\omega\in\Omega$ we have

$$\log\Lambda(t,\omega)\leqq\delta n\leqq\delta t,\quad\forall\, n\geqq N_\delta(\omega),\quad t\geqq n.$$

It follows that $\sup_{t\geq n}(1/t)\log\Lambda(t,\omega)\leq\delta$ holds for every $n\geq N_\delta(\omega)$, and thus also

$$\overline{\lim_{t\to\infty}}\frac{1}{t}\log\Lambda(t,\omega)\leq\delta \quad \text{for a.e.} \quad \omega\in\Omega.$$

The inequality of (9.24) is now a consequence of the arbitrariness of $\delta>0$.

10. Maximization of utility from both consumption and terminal wealth. Let us consider now an investor who derives utility both from "living well" (i.e., from consumption) and from "becoming rich" (i.e., from terminal wealth). His *expected total utility* is then

$$\begin{aligned} J(x;\pi,c) &\triangleq J_1(x;\pi,c)+J_2(x;\pi,c) \\ &= \mathbf{E}\int_0^T U_1(t,c(t))\,dt+\mathbf{E}U_2(T,X(T)), \end{aligned} \tag{10.1}$$

and he tries to maximize $J(x;\pi,c)$ over $\mathcal{A}_{1,2}(x)\triangleq\mathcal{A}_1(x)\cap\mathcal{A}_2(x)$:

$$V(x)\triangleq\sup_{(\pi,c)\in\mathcal{A}_{1,2}(x)}J(x;\pi,c). \tag{10.2}$$

Here again, U_1 and U_2 are utility functions in the sense of § 7.

In contrast to the problems of §§ 8 and 9, this one calls for balancing *competing objectives.* We shall show that the right compromise can be drawn in a very simple manner: at time $t=0$, the investor just divides his endowment x into two nonnegative parts x_1 and x_2, with $x_1+x_2=x$. For x_1, he solves the problem of § 8 (with utility U_1 from consumption); for x_2, he solves the problem of § 9 (with utility U_2 from terminal wealth). The superposition of his actions for these two problems will lead to the optimal policy for the problem of (10.2), provided x_1 and x_2 are chosen judiciously. We shall show exactly how this can be done (cf. (10.10) below).

For concreteness, we assume throughout this section that the value functions U_1, U_2 satisfy (7.2) and (7.3), as well as the requirements (8.8), (8.15) and (9.6), (9.11).

Let us start with an arbitrary pair $(\pi,x)\in\mathcal{A}_{1,2}(x)$ and define

$$x_1\triangleq\tilde{\mathbf{E}}\int_0^T\beta(t)c(t)\,dt, \qquad x_2\triangleq x-x_1. \tag{10.3}$$

Denoting by X the wealth process corresponding to this pair, we conclude from (4.3), (10.3) that

$$c\in\mathcal{D}(x_1), \qquad X(T)\in\mathcal{L}(x_2). \tag{10.4}$$

Theorem 8.1 gives us a pair $(\pi_1,c_1)\in\mathcal{A}_1(x_1)$ which is optimal for $V_1(x_1)$, with corresponding wealth process X_1 satisfying $X_1(T)=0$, almost surely. On the other hand, Theorem 9.1 provides a pair $(\pi_2,0)\in\mathcal{A}_2(x_2)$ which is optimal for $V_2(x_2)$, with corresponding wealth process X_2. If we define now

$$\tilde{\pi}\triangleq\pi_1+\pi_2, \quad \tilde{c}\triangleq c_1, \quad \text{and} \quad \tilde{X}\triangleq X_1+X_2 \tag{10.5}$$

and add (8.10), (9.8) memberwise, we obtain

$$\begin{aligned} \beta(t)\tilde{X}(t) &= \tilde{\mathbf{E}}\left[\int_t^T\beta(s)\tilde{c}(s)\,ds+\beta(T)\tilde{X}(T)\,\middle|\,\mathcal{F}_t\right] \\ &= x-\int_0^t\beta(s)\tilde{c}(s)\,ds+\int_0^t\beta(s)(\tilde{\pi}(s))^*\sigma(s)\,d\tilde{W}(s). \end{aligned} \tag{10.6}$$

In other words, $\tilde{X}$ *is the wealth process corresponding to the pair* $(\tilde{\pi},\tilde{c})\in\mathcal{A}_{1,2}(x)$.

But now recall (10.4), and add up (8.12), (9.10) memberwise to obtain

$$J(x;\pi,c)\leqq V_1(x_1)+V_2(x_2),$$

whence

$$V(x)\leqq V_*(x)\triangleq \max_{\substack{x_1\geqq 0,x_2\geqq 0\\ x_1+x_2=x}}[V_1(x_1)+V_2(x_2)]. \tag{10.7}$$

Therefore, the question is to find x_1, x_2 for which this maximum is achieved, because then the total expected utility corresponding to the pair $(\hat{\pi},\hat{c})$ of (10.5) will be *exactly* equal to $V_*(x)$; this will in turn imply

$$V(x)\equiv V_*(x) \tag{10.8}$$

from (10.7), and thus the above-mentioned pair will be shown to be *optimal* for the problem of (10.2).

But the maximization in (10.7) is easy: it amounts to selecting x_1, x_2 so that $V_1'(x_1)=V_2'(x_2)$ or, thanks to (8.17), (9.14): $\mathcal{Y}_1(x_1)=\mathcal{Y}_2(x_2)=\lambda\Leftrightarrow x_1=\mathcal{X}_1(\lambda), x_2=\mathcal{X}_2(\lambda)$. In other words, we find those values of x_1, x_2 for which the "marginal expected utilities" $V_1'(x_1)$, $V_2'(x_2)$ from the two individual optimization problems are identical.

The constant λ is determined uniquely as follows: we introduce the function

$$\mathcal{X}(y)\triangleq \mathcal{X}_1(y)+\mathcal{X}_2(y)=\mathbf{E}\left[\int_0^T \zeta(t)I_1(t,y\zeta(t))\,dt+\zeta(T)I_2(T,y\zeta(T))\right] \tag{10.9}$$

on $(0,\infty)$, which is continuous and strictly decreasing with $\mathcal{X}(0+)=\infty$, $\mathcal{X}(\infty)=0$. Let $\mathcal{Y}=\mathcal{X}^{-1}$ be the inverse of $\mathcal{X}$; then $\lambda=\mathcal{Y}(x)$, and the "optimal partition" of the initial wealth is given by

$$x_1=\mathcal{X}_1(\mathcal{Y}(x)),\qquad x_2=\mathcal{X}_2(\mathcal{Y}(x)). \tag{10.10}$$

If we also introduce the function

$$\begin{aligned} G(y)&\triangleq G_1(y)+G_2(y)\\ &=\mathbf{E}\left[\int_0^T U_1(t,I_1(t,y\zeta(t)))\,dt+U_2(T,I_2(T,y\zeta(T)))\right], \end{aligned} \tag{10.11}$$

which is continuous and decreasing on $(0,\infty)$, it is easy to see from (8.16), (9.12) that

$$V_*(x)=G(\mathcal{Y}(x)). \tag{10.12}$$

We have established the following result.

THEOREM 10.1. *Under the conditions of this section, the value function V of* (10.2) *is given as*

$$V=G\circ\mathcal{Y}. \tag{10.13}$$

For a fixed initial capital $x>0$, the optimal consumption rate process and the optimal level of terminal wealth are given by

$$\hat{c}(t)=I_1(t,\mathcal{Y}(x)\zeta(t)),\quad 0\leqq t\leqq T,\quad \text{and}\quad \hat{X}(T)\triangleq I_2(T,\mathcal{Y}(x)\zeta(T)), \tag{10.14}$$

respectively; there exists a portfolio process $\hat{\pi}$ such that $(\hat{\pi},\hat{c})$ is optimal in $\mathcal{A}_{1,2}(x)$ for (10.2), *and the corresponding wealth process $\hat{X}$ is given by*

$$\hat{X}(t)=\frac{1}{\beta(t)}\tilde{\mathbf{E}}\left[\int_t^T \beta(s)I_1(s,\mathcal{Y}(x)\zeta(s))\,ds+\beta(T)I_2(T,\mathcal{Y}(x)\zeta(T))\,\middle|\,\mathcal{F}_t\right] \tag{10.15}$$

almost surely, for every $0\leqq t\leqq T$.

Notice that the process M of (3.6), corresponding to the pair $(\hat{\pi}, \hat{c})$ of Theorem 10.1, takes the form

$$\hat{M}(t) = \tilde{\mathbf{E}}\left[\int_0^T \beta(s) I_1(s, \mathcal{Y}(x)\zeta(s))\, ds + \beta(T) I_2(T, \mathcal{Y}(x)\zeta(T)) \,\middle|\, \mathcal{F}_t\right];$$

in particular, it is a $\tilde{\mathbf{P}}$-martingale.

Example 10.2. In the case $U_1(t, c) = U_2(t, c) = \exp\{-\int_0^t \mu(s)\, ds\} \log c$, the functions of (10.9), (10.11), and (10.13) are given by

$$\mathcal{X}(y) = \frac{\alpha}{y}, \quad G(y) = -\alpha \cdot \log y + \delta, \quad 0 < y < \infty, \quad \text{and}$$

$$V(x) = \alpha \cdot \log\left(\frac{x}{\alpha}\right) + \delta, \qquad 0 < x < \infty$$

where $\alpha = \alpha_1 + \alpha_2$, $\delta = \delta + \delta_2$ in the notation of (8.19) and (9.16).

Example 10.3. In the case $U_1(t, c) = U_2(t, c) = -\exp\{-\int_0^t \mu(s)\, ds\}/c$, we obtain

$$\mathcal{X}(y) = \alpha y^{-1/2}, \quad G(y) = -\alpha y^{1/2}, \quad 0 < y < \infty, \quad \text{and}$$

$$V(x) = -\frac{\alpha^2}{x}, \qquad 0 < x < \infty$$

where $\alpha = \alpha_1 + \alpha_2$ in the notation of (8.21) and (9.21).

11. The case of constant coefficients. The theory developed in the last three sections, culminating with Theorem 10.1, provides a very precise characterization of the value function for the optimization problem (10.2) (cf. expression (10.13)), as well as explicit formulas for the optimal processes of consumption rate $\hat{c}$ and wealth $\hat{X}$ (in (10.14), (10.15), respectively). But for the optimal portfolio process $\hat{\pi}$, the "martingale methodology" that we have employed so far is able to ascertain only its existence (except in special cases, such as that of Example 9.3); in general, there is no constructive algorithm, or a useful characterization, that could lead to its computation.

Our intent in the present section is to improve this situation; we shall impose Draconian assumptions on the model, which will enable us in particular to obtain the optimal $\hat{\pi}$, $\hat{c}$ in a very explicit *feedback form on the current level of wealth* (cf. (11.23), (11.24)).

Specifically, we shall assume throughout this section that

$$r(t) \equiv r, \quad b(t) \equiv b, \quad \sigma(t) \equiv \sigma \quad \forall t \in [0, T] \tag{11.1}$$

for given $r \in \mathcal{R}$, $b \in \mathcal{R}^d$, and σ a nonsingular $(d \times d)$ matrix. We shall also assume separable utility functions, of the form $U_i(t, c) = e^{-\mu t} U_i(c)$, for $i = 1, 2$ and some real number $\mu \neq 0$. These assumptions will allow us to use "Markovian" methods, such as the Feynman-Kac representation of solutions to partial differential equations and the Hamilton-Jacobi-Bellman (HJB) equation of dynamic programming.

In order to make these methodologies available to us, we shall need a temporal as well as spatial parametrization; to wit, we write the analogues of the wealth equation (3.3) and of the value function (10.2) on the horizon $[t, T]$, for arbitrary $0 \leqq t \leqq T$, as

$$\begin{aligned} X(s) = x &+ \int_t^s (rX(u) - c(u))\, du + \int_t^s \pi^*(u)(b - r\mathbf{1})\, du \\ &+ \int_t^s \pi^*(u)\sigma\, dW(u), \qquad t \leqq s \leqq T \end{aligned} \tag{11.2}$$

and

$$V(t,x)=\sup_{(\pi,c)\in\mathcal{A}(t,x)}\mathbf{E}\left[\int_t^T e^{-\mu s}U_1(c(s))\,ds+e^{-\mu T}U_2(X(T))\right], \tag{11.3}$$

respectively. We shall also impose the purely technical assumptions

$$U_i(0)>-\infty,\quad \lim_{c\downarrow 0}\frac{(U_i'(c))^2}{U_i''(c)}\text{ exists},\quad \lim_{c\to\infty}\frac{(U_i'(c))^\alpha}{U_i''(c)}=0,\quad i=1,2 \tag{11.4}$$

for some $\alpha>2$. They will permit the analysis to go through conveniently, and include as special cases the so-called HARA (for Hyperbolic Absolute Risk Aversion) utility functions of the type $U(c)=(c+\eta)^\delta$; $0<\delta<1$, $\eta\geqq 0$. However, they are far from being the weakest possible conditions under which the fundamental results will hold.

By analogy with our previous analysis and notation, let us introduce the vector $\theta=\sigma^{-1}(b-r\mathbf{1})\in\mathcal{R}^d$, the processes

$$Z_s^{(t)}\triangleq\exp\{-\theta^*(W_s-W_t)-\tfrac{1}{2}\|\theta\|^2(s-t)\},\quad \phi_s^{(t)}\triangleq e^{(\mu-r)(s-t)}Z_s^{(t)},\quad\text{and}$$
$$Y_s^{(t,y)}\triangleq y\phi_s^{(t)},\qquad t\leqq s\leqq T,\quad 0<y<\infty,$$

as well as the functions $I_i=(U_i')^{-1}$, $i=1,2$ and

$$G(t,y)\triangleq\mathbf{E}\left[\int_t^T e^{-\mu(s-t)}U_1(I_1(Y_s^{(t,y)}))\,ds+e^{-\mu(T-t)}U_2(I_2(Y_T^{(t,y)}))\right], \tag{11.5}$$

$$\mathcal{X}(t,y)\triangleq\mathbf{E}\left[\int_t^T e^{-\mu(s-t)}\phi_s^{(t)}I_1(y\phi_s^{(t)})\,ds+e^{-\mu(T-t)}\phi_T^{(t)}I_2(y\phi_T^{(t)})\right], \tag{11.6}$$

$$S(t,y)\triangleq y\mathcal{X}(t,y)=\mathbf{E}\left[\int_t^T e^{-\mu(s-t)}Y_s^{(t,y)}I_1(Y_s^{(t,y)})\,ds+e^{-\mu(T-t)}Y_T^{(t,y)}I_2(Y_T^{(t,y)})\right] \tag{11.7}$$

for $(t,y)\in[0,T]\times(0,\infty)$. To avoid trivialities, we suppose $\theta\neq\mathbf{0}$; then for every $t\in[0,T)$ the function $\mathcal{X}(t,\cdot)$ is continuous and strictly decreasing on $(0,\infty)$, with $\mathcal{X}(t,0+)=\infty$ and $\mathcal{X}(t,\infty)=0$. We denote its inverse by $\mathcal{Y}(t,\cdot)$, i.e.,

$$\mathcal{Y}(t,\mathcal{X}(t,y))=y,\qquad 0\leqq t<T,\quad 0\leqq y\leqq\infty$$

and by analogy with the characterization (10.13) of Theorem 10.1 we have

$$V(t,x)=e^{-\mu t}G(t,\mathcal{Y}(t,x)),\qquad (t,x)\in[0,T]\times(0,\infty). \tag{11.8}$$

The point here is that we have reduced the study of the control problem (11.3) (or equivalently, of the nonlinear HJB equation (11.19) below, which is associated with it) to the study of the functions G, S of (11.5), (11.7); because, once these two are known, then $\mathcal{X}(t,y)$ is obtained straightaway as $y^{-1}S(t,y)$ and the value function V becomes available from (11.8). Now from the Feynman–Kac theorem, the functions G and S are characterized uniquely in terms of the Cauchy problems

$$\left(\frac{\partial}{\partial t}+L\right)G(t,y)+U_1(I_1(y))=0,\qquad (t,y)\in[0,T)\times(0,\infty) \tag{11.9}$$

$$G(T,y)=U_2(I_2(y)),\qquad y\in(0,\infty) \tag{11.10}$$

and

$$\left(\frac{\partial}{\partial t}+L\right)S(t,y)+yI_1(y)=0,\qquad (t,y)\in[0,T)\times(0,\infty) \tag{11.11}$$

$$S(T,y)=yI_2(y),\qquad y\in(0,\infty), \tag{11.12}$$

respectively, for the *linear* differential operator

$$L\varphi(t,y) \triangleq \frac{1}{2}\|\theta\|^2 y^2 \frac{\partial^2 \varphi(t,y)}{\partial y^2} + (\mu - r) y \frac{\partial \varphi(t,y)}{\partial y} - \mu \varphi(t,y).$$

Indeed, using the conditions (11.4) it can be shown that G, S satisfy growth conditions of the type

$$\max_{0 \leqq t \leqq T} |u(t,y)| \leqq K(1 + y^{\alpha} + y^{-\alpha}), \qquad 0 < y < \infty \tag{11.13}$$

for some positive constants α, K, and that among such functions they are the unique solutions of their respective Cauchy problems. *We shall show how to compute these solutions in closed form* (Proposition 11.1).

To this end, let us recall Remark 5.7 and observe that the unique solution to the auxiliary Cauchy problem

$$\left(\frac{\partial}{\partial t} + L\right) v(t,y;\xi) = 0, \qquad (t,y) \in [0,T) \times (0,\infty)$$

$$v(T,y;\xi) = (\xi - y)^+, \qquad y \in (0,\infty)$$

is given by the Black and Scholes-type formula for a "put" option (the right to *sell* one share of the stock at the pre-assigned price $\xi > 0$):

(11.14)

$$v(t,y;\xi) = \mathbf{E}[e^{-\mu(T-t)}(\xi - Y_T^{(t,y)})^+]$$

$$= \begin{cases} \xi e^{-\mu(T-t)} \Phi(-\nu_-(T-t,y;\xi)) - y e^{-r(T-t)} \Phi(-\nu_+(T-t,y;\xi)); & 0 \leqq t < T \\ (\xi - y)^+; & t = T \end{cases}$$

for every $(y,\xi) \in (0,\infty)^2$, where $\nu_{\pm}(t,y;\xi) \triangleq (1/\sqrt{2\gamma t})[\log(y/\xi) + t(\mu - r \pm \gamma)]$ and $\gamma \triangleq \|\theta\|^2/2$. Let us also introduce the functions

$$g(y) \triangleq \frac{U_1(I_1(y))}{\mu} - \frac{1}{\gamma(\lambda_+ - \lambda_-)} \left\{ \frac{y^{1+\lambda_+}}{1+\lambda_+} J_+(y) - \frac{y^{1+\lambda_-}}{1+\lambda_-} J_-(y) \right\}, \tag{11.15}$$

$$s(y) \triangleq \frac{y I_1(y)}{r} - \frac{1}{\gamma(\lambda_+ - \lambda_-)} \left\{ \frac{y^{1+\lambda_+}}{\lambda_+} J_+(y) - \frac{y^{1+\lambda_-}}{\lambda_-} J_-(y) \right\} \tag{11.16}$$

where $\lambda_+ > 0$ and $\lambda_- < 0$ are the roots of the quadratic equation $\gamma\lambda^2 - (r - \mu - \gamma)\lambda - r = 0$ and

$$J_+(y) = \int_0^{I_1(y)} (U_1'(c))^{-\lambda_+} \, dc, \qquad J_-(y) = \int_1^{I_1(y)} (U_1'(c))^{-\lambda_-} \, dc.$$

It is easy to verify that g, s solve the ordinary differential equations

$$Lg(y) + U_1(I_1(y)) = 0 \quad \text{and} \quad Ls(y) + y I_1(y) = 0,$$

respectively.

We can now put the various results together to arrive at the promised closed-form solutions.

PROPOSITION 11.1. *The functions G, S of* (11.5), (11.7) *have the stochastic representations*

$$G(t,y) = g(y) + \mathbf{E}[e^{-\mu(T-t)}\{U_2(I_2(Y_T^{(t,y)})) - g(Y_T^{(t,y)})\}],$$
$$S(t,y) = s(y) + \mathbf{E}[e^{-\mu(T-t)}\{Y_T^{(t,y)} I_2(Y_T^{(t,y)}) - s(Y_T^{(t,y)})\}],$$

which lead to the closed-form expressions

$$G(t,y) = g(y) + \left(U_2(0) - \frac{U_1(0)}{\mu}\right) e^{-\mu(T-t)} + \int_0^\infty (U_2(I_2(\xi)) - g(\xi))'' v(t,y;\xi)\, d\xi, \tag{11.17}$$

$$S(t,y) = s(y) + \int_0^\infty (\xi I_2(\xi) - s(\xi))'' v(t,y;\xi)\, d\xi. \tag{11.18}$$

Moreover, the function $V:[0,T]\times[0,\infty)\to\mathcal{R}$ *of* (11.8) *satisfies the* HJB *Equation of Dynamic Programming*

$$\max_{\substack{c\geqq 0\\ \pi\in\mathcal{R}^d}} [\tfrac{1}{2}\|\pi^*\sigma\|^2 V_{xx}(t,x) + \{(rx-c) + \pi^*(b-r\mathbf{1})\} V_x(t,x) + e^{-\mu t} U_1(c)] = -V_t(t,x) \quad \text{in } [0,T)\times(0,\infty) \tag{11.19}$$

and the terminal-boundary conditions

$$V(T,x) = e^{-\mu T} U_2(x), \qquad 0 \leqq x < \infty, \tag{11.20}$$

$$V(t,0) = \left(U_2(0) - \frac{U_1(0)}{\mu}\right) e^{-\mu T} + \frac{U_1(0)}{\mu} e^{-\mu t}, \qquad 0 \leqq t \leqq T, \tag{11.21}$$

respectively.

It is noteworthy that we have obtained a closed-form solution for the *nonlinear* HJB equation (11.19), by solving instead the two *linear* equations (11.9), (11.11) subject to the appropriate terminal and growth conditions, and then performing the composition (11.8).

The maximizations over $c\geqq 0$, $\pi\in\mathcal{R}^d$ in (11.19) are achieved by

$$\hat{c} = I_1(\mathcal{Y}(t,x)), \qquad \hat{\pi} = -(\sigma\sigma^*)^{-1}(b-r\mathbf{1})\frac{\mathcal{Y}(t,x)}{\mathcal{Y}_x(t,x)}.$$

This suggests that we should be able to justify similar *feedback form expressions* (on the current level of wealth) *for the optimal consumption rate and portfolio processes.* We choose to do this by studying directly the optimal wealth process.

PROPOSITION 11.2. *The optimal wealth process* $X^{(t,x)}$ *for the problem* (11.3) *is given by*

$$X^{(t,x)}(s) = \mathcal{X}(s, \eta_s^{(t,x)}), \qquad t \leqq s \leqq T, \tag{11.22}$$

where $\eta_s^{(t,x)} \triangleq \mathcal{Y}(t,x)\phi_s^{(t)} = Y_s^{(t,\mathcal{Y}(t,x))}$. *In terms of* $X^{(t,x)}$, *the optimal pair* $(\pi^{(t,x)}, c^{(t,x)})$ *can be expressed as*

$$c^{(t,x)}(s) = I_1(\mathcal{Y}(s, X^{(t,x)}(s))), \tag{11.23}$$

$$\pi^{(t,x)}(s) = -(\sigma\sigma^*)^{-1}(b-r\mathbf{1})\frac{\mathcal{Y}(s, X^{(t,x)}(s))}{\mathcal{Y}_x(s, X^{(t,x)}(s))} \tag{11.24}$$

for $t \leqq s \leqq T$.

Proof. By analogy with (10.15), the optimal wealth process is almost surely given by

$$X^{(t,x)}(s) = \mathbf{E}\left[\int_s^T e^{-r(\theta-s)} Z_\theta^{(s)} I_1(\eta_\theta^{(t,x)})\, d\theta + e^{-r(T-s)} Z_T^{(s)} I_2(\eta_T^{(t,x)}) \,\middle|\, \mathcal{F}_s\right]$$

$$= \frac{1}{Y_s^{(t,y)}} \mathbf{E}\left[\int_s^T e^{-\mu(\theta-s)} Y_\theta^{(t,y)} I_1(Y_\theta^{(t,y)})\, d\theta + e^{-\mu(T-t)} Y_T^{(t,y)} I_2(Y_T^{(t,y)}) \,\middle|\, \mathcal{F}_s\right]$$

$$= \frac{S(s, Y_s^{(t,y)})}{Y_s^{(t,y)}}$$

with $y = \mathcal{Y}(t, x)$, for every $s \in [t, T]$; but this is (11.22). Now it is easily seen from (3.10) that $\eta^{(t,x)}$ satisfies the linear stochastic equation

$$d\eta_s^{(t,x)} = \eta_s^{(t,x)}[(\mu - r)\, ds - \theta^*\, dW(s)], \qquad \eta_t^{(t,x)} = x. \tag{11.25}$$

On the other hand, by substituting $S(t, y) = y\mathcal{X}(t, y)$ into (11.11), we arrive at the linear parabolic equation

$$\mathcal{X}_t + \gamma y^2 \mathcal{X}_{yy} + (\mu - r + 2\gamma) y \mathcal{X}_y - r\mathcal{X} + I_1(y) = 0, \qquad 0 \leqq t < T, \quad 0 < y < \infty \tag{11.26}$$

for $\mathcal{X}(t, y)$. An application of Itô's rule to (11.22), in conjunction with (11.25) and (11.26), leads to

$$dX^{(t,x)}(s) = (rX^{(t,x)}(s) - c^{(t,x)}(s))\, ds + (\pi^{(t,x)}(s))^*[(b - r\mathbf{1})\, ds + \sigma\, dW(s)]$$

in the notation of (11.23), (11.24). But a comparison of this equation with (3.3) shows that $X^{(t,x)}$ is the wealth associated with the pair $(\pi^{(t,x)}, c^{(t,x)})$. □

Example 11.3. In the special case $U_1(c) = U_2(c) = c^\delta$ for some $0 < \delta < 1$, we have

$$G(t, y) = p(t)\left(\frac{y}{\delta}\right)^{\delta/(\delta-1)}, \quad S(t, y) = \delta G(t, y), \quad \mathcal{X}(t, y) = p(t)\left(\frac{y}{\delta}\right)^{1/(\delta-1)}$$

and

$$V(t, x) = e^{-\mu t}(p(t))^{1-\delta} x^\delta,$$

as well as

$$c^{(t,x)}(s) = \frac{X^{(t,x)}(s)}{p(s)}, \quad \pi^{(t,x)}(s) = (\sigma\sigma^*)^{-1}(b - r\mathbf{1})\frac{X^{(t,x)}(s)}{1-\delta}, \quad t \leqq s \leqq T,$$

where

$$p(t) = \begin{cases} (1/k)[1 - e^{-k(T-t)}] + e^{-k(T-t)}; & k \neq 0 \\ 1 + T - t & k = 0 \end{cases},$$

$$k = \frac{1}{1-\delta}\left(\mu - r\delta - \frac{\gamma\delta}{1-\delta}\right).$$

12. An equilibrium model. Let us consider in this final section an *economy* that consists of the following:

(i) The same financial market as in § 2;

(ii) A single consumption good or "commodity," traded at the spot price $\psi = \{\psi(t);\ 0 \leqq t \leqq T\}$; and

(iii) A finite number n of economic agents (small investors). Each one receives an exogenous *endowment* at the rate $\varepsilon_k = \{\varepsilon_k(t);\ 0 \leqq t \leqq T\}$, denominated in units of the commodity; he can either consume this endowment or turn it into cash, and invest the proceeds in the financial market. Each agent also has a *utility function* U_k, and attempts to maximize his expected total utility from consumption (as in § 8).

The *equilibrium problem* for such an economy is to determine a spot price ψ, so that the markets clear when each agent behaves optimally and the commodity is traded at the price ψ. We shall show that the methodology of § 8 is ideally suited to handle this question.

We shall assume throughout this section that the processes ψ and $\varepsilon_1, \cdots, \varepsilon_n$ are positive and progressively measurable with respect to $\{\mathcal{F}_t\}$. On the other hand, both the "deflated spot price" $\zeta(t)\psi(t)$ and the "aggregate endowment rate"

$$\varepsilon(t) = \sum_{k=1}^{n} \varepsilon_k(t), \qquad 0 \leqq t \leqq T \tag{12.1}$$

processes, will be assumed to take values in intervals of the form $[\delta, \Delta]$, for finite constants $\Delta > \delta > 0$.

For a given spot price process ψ, the kth agent has at his disposal the choice of a portfolio process $\pi_k(t) = (\pi_{k1}(t), \cdots, \pi_{kd}(t))^*$ and a consumption rate process $c_k(t)$, $0 \leqq t \leqq T$, as in Definitions 3.1 and 3.2 (except that (3.2) is now replaced by $\int_0^T \psi(t)c_k(t)\,dt < \infty$, almost surely). For every such pair (π_k, c_k), the corresponding *wealth process* X_k satisfies, by analogy with (3.3), the equation

$$\begin{aligned} dX_k(t) = r(t)X_k(t)\,dt + \psi(t)[\varepsilon_k(t) - c_k(t)]\,dt + \pi_k^*(t)[b(t) - r(t)\mathbf{1}]\,dt \\ + \pi_k^*(t)\sigma(t)\,dW(t). \end{aligned}$$

In terms of the $\tilde{\mathbf{P}}$-Brownian motion $\tilde{W}$ of (2.9), the solution is given by

$$\beta(t)X_k(t) = \int_0^t \beta(s)\psi(s)[\varepsilon_k(s) - c_k(s)]\,ds + \int_0^t \beta(s)\pi_k^*(s)\sigma(s)\,d\tilde{W}(s). \tag{12.2}$$

The kth agent's *optimization problem* is to maximize the expected total utility

$$\mathbf{E}\int_0^T U_k(t, c_k(t))\,dt \tag{12.3}$$

from consumption, over all admissible pairs (π_k, c_k)—as in Definition 4.1—that satisfy

$$\mathbf{E}\int_0^T U_k^-(t, c_k(t))\,dt < \infty. \tag{12.4}$$

Let us denote by $(\hat{\pi}_k, \hat{c}_k)$ the optimal pair for this problem, and by $\hat{X}_k$ the associated wealth process.

We are now in a position to define the notion of equilibrium for the economy.

DEFINITION 12.1. A spot price process ψ is called an *equilibrium spot price process*, if we have the following:

(a) Clearing of the commodity market, i.e.,

$$\sum_{k=1}^{n} \hat{c}_k(t) = \varepsilon(t), \qquad \forall\, 0 \leqq t \leqq T, \qquad \text{and} \tag{12.5}$$

(b) Clearing of the financial markets, i.e.,

$$\sum_{k=1}^{n} \hat{\pi}_{ki}(t) = 0, \quad \lambda \times \mathbf{P}\text{-a.e. on } [0, T] \times \Omega, \qquad \forall\, i = 1, \cdots, d \tag{12.6}$$

$$\sum_{k=1}^{n} \hat{X}_k(t) = 0, \qquad \forall\, 0 \leqq t \leqq T \tag{12.7}$$

almost surely.

For a given ψ, let us try to solve the kth agent's optimization problem. First, here is an analogue of Theorem 4.4 and Proposition 4.5.

THEOREM 12.2. *Every admissible pair (π_k, c_k) for the kth agent satisfies*

$$\mathbf{E}\int_0^T \zeta(t)\psi(t)c_k(t)\,dt \leqq \mathbf{E}\int_0^T \zeta(t)\psi(t)\varepsilon_k(t)\,dt. \tag{12.8}$$

Conversely, if a consumption rate process c_k satisfies (12.8), *there exists a portfolio π_k such that the pair (π_k, c_k) is admissible for the kth agent. In particular, if* (12.8) *is satisfied as an equality, then the portfolio π_k is unique up to equivalence and the corresponding wealth process X_k is given by*

$$\beta(t)X_k(t) = \tilde{\mathbf{E}}\left[\int_t^T \beta(s)(c_k(s)-\varepsilon_k(s))\,ds \,\middle|\, \mathcal{F}_t\right], \qquad 0 \leqq t \leqq T. \tag{12.9}$$

Proof. For every portfolio/consumption process pair (π_k, c_k), the analogues

$$M_k(t) = \beta(t)X_k(t) + \int_0^t \beta(s)\psi(s)[c_k(s)-\varepsilon_k(s)]\,ds, \tag{12.10}$$

$$N_k(t) = \zeta(t)X_k(t) + \int_0^t \zeta(s)\psi(s)[c_k(s)-\varepsilon_k(s)]\,ds \tag{12.11}$$

of the processes in (3.6), (3.9) are continuous, local martingales under $\tilde{\mathbf{P}}$ and $\mathbf{P}$, respectively. Now if the pair (π_k, c_k) is admissible, N_k is bounded from below by a $\mathbf{P}$-integrable random variable, and is thus a $\mathbf{P}$-supermartingale (which implies that M_k is a $\tilde{\mathbf{P}}$-supermartingale); (12.8) follows from this property, in conjunction with $X_k(T) \geqq 0$, almost surely.

Conversely, for every consumption rate process c_k that satisfies (12.8), we introduce the random variable

$$D_k = \int_0^T \beta(s)\psi(s)[\varepsilon_k(s)-c_k(s)]\,ds; \tag{12.12}$$

the condition (12.8) amounts to $\tilde{\mathbf{E}}D_k \geqq 0$, and the martingale

$$u_k(t) \triangleq \tilde{\mathbf{E}}D_k - \tilde{\mathbf{E}}(D_k|\mathcal{F}_t), \qquad 0 \leqq t \leqq T \tag{12.13}$$

is representable as in (4.9):

$$u_k(t) = \int_0^t \beta(s)\pi_k^*(s)\sigma(s)\,d\tilde{W}(s), \tag{12.14}$$

for a suitable portfolio process π_k. From (12.2) and (3.8), the wealth process X_k corresponding to (π_k, c_k) is given as

$$\begin{aligned} X_k(t) &= \frac{1}{\beta(t)}\left[\int_0^t \beta(s)\psi(s)\{\varepsilon_k(s)-c_k(s)\}\,ds + u_k(t)\right] \\ &= \frac{1}{\zeta(t)}\left[Z(t)\tilde{\mathbf{E}}D_k - \mathbf{E}\left(\int_t^T \zeta(s)\psi(s)\{\varepsilon_k(s)-c_k(s)\}\,ds \,\middle|\, \mathcal{F}_t\right)\right], \end{aligned} \tag{12.15}$$

almost surely. Both requirements of (4.1) for admissibility follow easily from this last representation.

The remaining claims follow as in the proof of Proposition 4.5. □

According to Theorem 12.2, the kth agent's optimization problem is reformulated as follows: *to maximize* $\mathbf{E}\int_0^T U_k(t, c_k(t))\,dt$, *over consumption rate processes c_k that*

satisfy (12.8) *and* (12.4). But the solution to this problem is known from § 8: the optimal consumption rate process is of the form

$$\hat{c}_k(t) = I_k(t, y_k\zeta(t)\psi(t)), \qquad 0 \leqq t \leqq T, \quad k \in \{1, \cdots, n\} \tag{12.16}$$

where y_k is the unique number in $(0, \infty)$ that satisfies

$$\mathbf{E}\int_0^T \zeta(t)\psi(t)I_k(t, y_k\zeta(t)\psi(t))\,dt = \mathbf{E}\int_0^T \zeta(t)\psi(t)\varepsilon_k(t)\,dt, \qquad k = 1, \cdots, n. \tag{12.17}$$

In particular, $\hat{c}_k$ satisfies (12.8) as an equality.

The outstanding question now is whether we can find a spot price process ψ for which (12.5)–(12.7) are satisfied.

PROPOSITION 12.3. *Let ψ be an equilibrium spot price process, and let the vector $Y = (y_1, \cdots, y_n) \in (0, \infty)^n$ be defined in terms of ψ by* (12.17). *Then ψ and Y must satisfy*

$$\sum_{k=1}^{n} I_k(t, y_k\zeta(t)\psi(t)) = \varepsilon(t), \qquad 0 \leqq t \leqq T. \tag{12.18}$$

Conversely, suppose that $Y \in (0, \infty)^n$ and the spot price process ψ are such that (12.17), (12.18) *are satisfied; then ψ is an equilibrium spot price process.*

Proof. The first claim follows directly, by substituting the expressions (12.16) into (12.5). For the second claim notice that, under the spot price ψ in question, the optimal consumption processes $\{\hat{c}_k\}_{k=1}^n$ are still given by (12.16) and satisfy (12.5); letting $\hat{D}_k$, $\hat{u}_k$ and $\hat{\pi}_k$, $\hat{X}_k$ be the corresponding quantities in (12.12), (12.13), and (12.14), (12.9), respectively, we obtain with the help of (12.5), (12.1): $\sum_{k=1}^n \hat{X}_k(t) \equiv 0, \sum_{k=1}^n \hat{D}_k = 0$, and $\sum_{k=1}^n \hat{u}_k(t) \equiv 0$, almost surely. From this last identity and (12.14), we conclude that (12.6) holds. □

A further reduction in the characterization of equilibrium is obtained by introducing the function

$$I(t, h; \Lambda) \triangleq \sum_{k=1}^{n} I_k(t, h\lambda_k^{-1}), \qquad (t, h) \in [0, T] \times (0, \infty) \tag{12.19}$$

for every $\Lambda = (\lambda_1, \cdots, \lambda_n) \in (0, \infty)^n$, and denoting by $H(t, \cdot\,; \Lambda)$ the inverse of the strictly decreasing mapping $I(t, \cdot\,; \Lambda)\colon [0, \infty] \xrightarrow{\text{onto}} [0, \infty]$, with fixed $t \in [0, T]$ and $\Lambda \in (0, \infty)^n$. The function H enjoys the *positive homogeneity property*

$$H(t, c; \rho\Lambda) = \rho H(t, c; \Lambda), \quad \forall \rho > 0 \tag{12.20}$$

and in terms of H the equations (12.18), (12.17) are rewritten as

$$\psi(t) \equiv \psi(t; \Lambda) = \frac{1}{\zeta(t)} H(t, \varepsilon(t); \Lambda), \qquad 0 \leqq t \leqq T, \tag{12.21}$$

$$\begin{aligned} &\mathbf{E}\int_0^T H(t, \varepsilon(t); \Lambda) I_k\left(t, \frac{1}{\lambda_k} H(t, \varepsilon(t); \Lambda)\right) dt \\ &\qquad = \mathbf{E}\int_0^T H(t, \varepsilon(t); \Lambda)\varepsilon_k(t)\,dt, \qquad k = 1, \cdots, n, \end{aligned} \tag{12.22}$$

respectively, with the identification $\Lambda = (\lambda_1, \cdots, \lambda_n) = (1/y_1, \cdots, 1/y_n) \in (0, \infty)^n$.

We conclude from this analysis that *the search for equilibrium has been reduced to the search for a vector $\Lambda \in (0, \infty)^n$ which satisfies* (12.22); the corresponding equilibrium

spot price and optimal consumption rate processes would then be given by (12.21) and

$$\hat{c}_k(t) \equiv \hat{c}_k(t;\Lambda) = I_k\left(t, \frac{1}{\lambda_k} H(t, \varepsilon(t);\Lambda)\right), \qquad 0 \leqq t \leqq T, \quad k = 1, \cdots, n, \tag{12.23}$$

respectively (cf. (12.16)).

Now it is seen from (12.20), (12.22) that if $\Lambda \in (0,\infty)^n$ is a solution of (12.22), then the entire ray $\{\rho\Lambda;\ \rho \in (0,\infty)\}$ is a family of solutions. The following result provides a sufficient condition, under which there is only one such ray.

THEOREM 12.4. *Suppose that, for every* $(t,k) \in [0,T] \times \{1, \cdots, n\}$,

$$c \mapsto cU_k'(t,c) \quad \textit{is a nondecreasing function.} \tag{12.24}$$

Then there exists a vector $\Lambda \in (0,\infty)^n$ *satisfying* (12.22); *if* Λ_1, Λ_2 *are two such vectors, then there exists a* $\gamma > 0$ *such that*

$$\Lambda_1 = \gamma\Lambda_2, \qquad \psi(\cdot;\Lambda_1) = \gamma\psi(\cdot;\Lambda_2) \tag{12.25}$$

$$\hat{c}_k(\cdot;\Lambda_1) = \hat{c}_k(\cdot;\Lambda_2), \qquad k = 1, \cdots, n. \tag{12.26}$$

In other words, equilibrium spot prices can be determined only up to a multiplicative constant, since there can always be a re-valuation of currency; this is not going to affect, however, the way in which *real* wealth, measured in units of optimal consumption for the commodity, will be distributed in equilibrium among the agents.

Example 12.5. Suppose that all the agents have the same utility function $U_k(t,c) = c^\delta \exp\{-\int_0^t \mu(s)\,ds\}$, with $0 < \delta < 1$ and μ as in Example 8.3. Then

$$\lambda_k = \left(\frac{\mathbf{E}\int_0^T \exp\left(-\int_0^t \mu(s)\,ds\right)\varepsilon_k(t)(\varepsilon(t))^{\delta-1}\,dt}{\mathbf{E}\int_0^T \exp\left(-\int_0^t \mu(s)\,ds\right)(\varepsilon(t))^{\delta}\,dt}\right)^{1-\delta}, \qquad k = 1, \cdots, n \tag{12.27}$$

gives the unique solution of (12.22) subject to $\sum_{k=1}^n \lambda_k^{1/(1-\delta)} = 1$; the corresponding processes of (12.21), (12.23) are

$$\psi(t) = \frac{\text{const.} \exp\left(-\int_0^t \mu(s)\,ds\right)}{\zeta(t)(\varepsilon(t))^{1-\delta}}, \qquad \hat{c}_k(t) = \lambda_k^{1/(1-\delta)}\varepsilon(t). \tag{12.28}$$

Example 12.6. If $U_k(t,c) = \exp\{-\int_0^t \mu(s)\,ds\}\log c$, for every $k \in \{1, \cdots, n\}$, we obtain the same results as in (12.27), (12.28) but with $\delta = 0$.

Example 12.7. Suppose that $U_k(t,c) = -\exp\{-\int_0^t \mu(s)\,ds\}/c$ holds for every $k \in \{1, \cdots, n\}$. Then the unique solution of (12.22) subject to $\sum_{k=1}^n \lambda_k^{1/2} = 1$ is given by

$$\lambda_k = \left(\frac{\mathbf{E}\int_0^T \exp\left(-\int_0^t \mu(s)\,ds\right)(\varepsilon(t))^{-2}\varepsilon_k(t)\,dt}{\mathbf{E}\int_0^T \exp\left(-\int_0^t \mu(s)\,ds\right)(\varepsilon(t))^{-1}\,dt}\right)^2, \qquad k = 1, \cdots, n.$$

The equilibrium spot price and optimal consumption rate processes are then given as

$$\psi(t) = \frac{\text{const.} \exp\left(-\int_0^t \mu(s)\,ds\right)}{\zeta(t)(\varepsilon(t))^2}, \qquad \hat{c}_k(t) = \lambda_k^{1/2}\varepsilon(t).$$

It should be noted that the condition (12.24) is satisfied by the utility functions of Examples 12.5 and 12.6, but not by that of Example 12.7 (in fact, in this latter example, the function $c \mapsto cU_k'(t,c)$ is strictly decreasing). Thus, the condition (12.24) is far from being necessary for the validity of the results of Theorem 12.4.

We send the interested reader to Karatzas, Lakner, Lehoczky, and Shreve (1988) for the proof of Theorem 12.4 and suggest the relaxation (or removal) of condition (12.24) as an interesting open problem.

13. Notes.

Section 1. The theory of continuous trading is a specialized but important topic in financial economics; see the recent books by Ingersoll (1987) and Duffie (1988) for a broad and exhaustive overview of the theory of financial decision making. The book by Malliaris and Brock (1982) is a good survey of stochastic models in economics and finance; for the early work on this subject, see the articles in the volume edited by Cootner (1964), in particular the papers "Brownian Motion in the Stock Market" and "Periodic Structure in the Brownian Motion of Stock Prices," by M. Osborne.

Sections 2, 3. The idea of introducing a probability measure, under which the discounted stock prices of (2.12) are martingales, is due to Harrison and Kreps (1979) and Harrison and Pliska (1981), (1983). The model that we have adopted, with the particular nondegeneracy condition (2.3), is due to Bensoussan (1984).

Condition (2.3) is essential in our development; by contrast, the boundedness of the process $r(\cdot)$, $b(\cdot)$ has been imposed only for simplicity. If we assume that these processes are just square-integrable almost surely on $[0, T]$, then the entire analysis on a finite horizon $[0, T]$ goes through with minor changes, provided that the process Z of (2.7)—always a supermartingale under the very weak almost sure condition $\int_0^T \|\theta(t)\|^2\, dt < \infty$—is actually a *martingale.* A sufficient condition for this, due to Novikov, is $E[\exp\{\frac{1}{2}\int_0^T \|\theta(t)\|^2\, dt\}] < \infty$ (cf. Karatzas and Shreve (1987, Prop. 3.5.12). We only have to replace the condition (5.1) by (5.2), and (6.3) by the requirement $\tilde{E}[\sup_{0\leq t\leq T} f(t)\beta(t)] < \infty$.

For results with "incomplete market models" (i.e., with more sources of uncertainty than stocks in the market model), see Föllmer and Sondermann (1986), Schweizer (1988), Pagès (1987), He and Pearson (1988), and Pagès (1989); consult also Karatzas, Lehoczky, Shreve, and Xu (1989).

Section 4. The material here is drawn from Karatzas and Shreve (1987, § 5.8) and Karatzas, Lehoczky, and Shreve (1987). The terminology "attainable levels of terminal wealth" is due to Pliska (1986), who has a result similar to Proposition 4.7.

The existence of an equivalent probability measure under which the discounted prices are martingales (an "equivalent martingale measure" as it is sometimes called), implies the absence of arbitrage opportunities (Remark 4.8). This property holds for very general price processes. The converse, i.e., the existence of an equivalent martingale measure in the absence of arbitrage opportunities, is known to hold in discrete time (cf. Taqqu and Willinger (1987) for finite probability spaces, and the very recent work by Dalang, Morton, and Willinger (1988) for arbitrary probability spaces). As far as we know, the question is open for general, continuous-time price processes.

Section 5. We follow Karatzas and Shreve (1987). Example 5.6 is adapted from Harrison and Pliska (1981). For comprehensive accounts of option pricing, see Samuelson (1973), Merton (1973), Smith (1976), and Ingersoll (1987). For a model of option pricing, in which the borrowing rate is higher than the interest rate of the bond, see Barron and Jensen (1988).

Section 6. Section 6 is adapted from Karatzas (1988); see also Bensoussan (1984), for a different approach to the stopping problem. For more recent work, in general semimartingale models, see the doctoral dissertation of Schweizer (1988).

Section 7. The assumption $U'(t, 0+) = \infty$ is made only for simplicity; for full treatments of the *consumption/investment* and *equilibrium* problems that do not rely on this assumption, see Karatzas, Lehoczky, and Shreve (1987) and (1988), respectively.

Sections 8–11. Sections 8–11 come essentially from Karatzas, Lehoczky, and Shreve (1987) (with the exception of §§ 9.3, 9.4, 9.6); that article should be consulted for some details which are only sketched here. The model with constant coefficients and utility from consumption was introduced by Samuelson (1969) and Merton (1969), (1971) for utility functions of the HARA class; it was studied in great detail by Karatzas et al. (1986) for general utility functions, using the HJB equation of dynamic programming and allowing for general patterns of behaviour upon *bankruptcy.*

Related results have been obtained independently by Cox and Huang (1986), (1987). For models with *constraints* (on borrowing, short-selling, etc.) see the recent works by Xu (1989), Zariphopoulou (1989), Grossman and Vila (1988), He and Pearson (1988).

Recent work by Davis and Norman (1988) treats a model with constant coefficients, one stock, and utility $U(t, c) = e^{-\mu t} c^{\delta}$, $0 < \delta < 1$ from consumption, but with *costs of transaction* between the two assets (see also Taksar et al. (1988), Eastham and Hastings (1988), Leland (1985)). On the other hand, Pontier and Picqué (1988) discuss the consumption/investment problem in a market model with $d = 2$, and stock prices driven by independent Brownian and Poisson processes. For a model with consumption in several goods with quite arbitrary prices, see Lakner (1989).

For maximization of the growth rate from investment, in discrete-time settings, see the articles by Cover (1984), (1988), and Algoet and Cover (1988), as well as Breiman (1961), Hakansson and Lin (1970), Hakansson (1971), Latané (1959), and Thorp (1971).

Much of the analysis in § 11 goes through if $r(\cdot), b(\cdot), \sigma(\cdot)$ are deterministic functions of time. We can also handle models in which the interest rate $r(\cdot)$ and the stock prices $\underset{\sim}{P}(t) = (P_1(t), \cdots, P_d(t))$ are quite general Markov diffusion processes driven by the Brownian motion W; the optimal portfolio/consumption policies can then be obtained in feedback form on the current level of the "extended state" $(X(t), r(t), \underset{\sim}{P}(t))$.

As the referee points out, the objective functions of (8.1), (10.1) mandate that utility at time t is derived from the level of consumption at that time only, for every $t \in (0, T)$; this excludes, for instance, situations like the "ratchet effect" (according to which the utility derived from consumption at one time is influenced by consumption levels at earlier times in such a way that a rise of consumption levels produces high utility but a decline produces low utility). A recent reference on such intertemporal dependence on preferences is Sundaresan (1985).

Section 12. Section 12 is taken from Karatzas et al. (1988). See also Duffie (1986), Duffie and Huang (1985), (1987), Duffie and Zame (1988) and Huang (1987) for related equilibrium models, as well as Cox, Ingersoll, and Ross (1985).

Acknowledgments. This paper is an expanded version of my notes entitled "Applications of Stochastic Calculus in Financial Economics," for lectures delivered at the Systems Research Center, University of Maryland, July 20–31, 1987. These notes are scheduled to appear in the series Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, New York.

I wish to thank Professor Harold Kushner for the invitation to transform those lecture notes into a full-fledged survey paper, and to express my appreciation to the referees for their careful reading of the manuscript and for their helpful comments and suggestions. Thanks are also due to Professor Herbert Robbins, for bringing to my attention the articles of Algoet and Cover (1988) and Cover (1984), (1988), which prompted the investigation that led to the results of § 9.6.

I am indebted to Mr. X. X. Xue for ideas and arguments that led to the proof of uniqueness in Propositions 4.5 and 4.7, and to Dr. Walter Willinger for helpful discussions and for bringing the results of his work (Dalang, Morton, and Willinger (1988)) to my attention.

Dedication. This article is dedicated to Vic Beneš with affection and respect.

REFERENCES

P. H. ALGOET AND T. M. COVER (1988), *Asymptotic optimality and asymptotic equipartition properties of log-optimum investment*, Ann. Probab., 16, pp. 876–898.

E. N. BARRON AND R. JENSEN (1988), *A stochastic control approach to the pricing of options*, Math. Oper. Res., to appear.

A. BENSOUSSAN (1984), *On the theory of option pricing*, Acta Appl. Math., 2, pp. 139–158.

J. M. BISMUT AND B. SKALLI (1977), *Temps d'arrêt optimal, théorie générale de processus et processus de Markov*, Z. Wahrsch. Verw. Gebiete, 39, pp. 301–313.

F. BLACK AND M. SCHOLES (1973), *The pricing of options and corporate liabilities*, J. Political Economy, 81, pp. 637–659.

L. BREIMAN (1961), *Optimal gambling systems for favorable games*, in Proc. 4th Berkeley Symposium Math. Statist. and Probability I, University of California Press, Berkeley, CA, pp. 65–78.

P. H. COOTNER, ED. (1964), *The Random Character of Stock Market Prices*, MIT Press, Cambridge, MA.

T. M. COVER (1984), *An algorithm for maximizing expected log investment return*, IEEE Trans. Inform. Theory, 30, pp. 369–373.

——— (1988), *Universal portfolios*, Technical Report 66, Department of Statistics, Stanford University, Stanford, CA.

J. C. COX, J. E. INGERSOLL, AND S. ROSS (1985), *An intertemporal general equilibrium model of asset prices*, Econometrica, 53, pp. 363–384.

J. C. COX AND C. F. HUANG (1986), *A variational problem arising in financial economics with an application to a portfolio turnpike theorem*, preprint.

——— (1987), *Optimal consumption and investment policies when asset prices follow a diffusion process*, preprint.

R. C. DALANG, A. MORTON, AND W. WILLINGER (1988), *Existence and uniqueness of equivalent martingale measures for vector-valued processes in discrete and finite time*, preprint; Stochastics, to appear.

M. H. A. DAVIS AND A. R. NORMAN (1988), *Portfolio selection with transaction costs*, Math. Oper. Res., to appear.

D. DUFFIE (1986), *Stochastic equilibria: existence, spanning number, and the "no expected financial gain from trade" hypothesis*, Econometrica, 54, pp. 1161–1183.

D. DUFFIE AND C. F. HUANG (1985), *Implementing Arrow-Debreu equilibria by continuous trading of few long-lived securities*, Econometrica, 53, pp. 1337–1356.

——— (1987), *Stochastic production-exchange equilibria*, preprint.

D. DUFFIE AND W. ZAME (1988), *The consumption-based capital asset pricing model*, preprint.

N. EL KAROUI (1981), *Les aspects probabilistes du contrôle stochastique*, Lecture Notes in Mathematics, 876, Springer-Verlag, Berlin, pp. 73–238.

A. G. FAKEEV (1970), *Optimal stopping rules for processes with continuous parameter*, Theory Probab. Appl., 15, pp. 324–331.

——— (1971), *Optimal stopping of a Markov process*, Theory Probab. Appl., 16, pp. 694–696.

I. V. GIRSANOV (1960), *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probab. Appl., 5, pp. 285–301.

N. HAKANSSON (1971), *Capital growth and the mean-variance approach to portfolio selection*, J. Financial Quantit. Anal., VI, pp. 517–557.

N. HAKANSSON AND T. LIU (1970), *Optimal growth portfolios when yields are serially correlated*, Rev. Econom. Statist., 52, pp. 385–394.

J. M. HARRISON AND D. M. KREPS (1979), *Martingales and arbitrage in multiperiod security markets*, J. Econom. Theory, 20, pp. 381–408.

J. M. HARRISON AND S. R. PLISKA (1981), *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11, pp. 215–260.

——— (1983), *A stochastic calculus model of continuous trading: complete markets*, Stochastic Process. Appl., 15, pp. 313–316.

C. F. HUANG (1987), *An intertemporal general equilibrium asset pricing model: the case of diffusion information*, Econometrica, 55, pp. 117–142.

N. IKEDA AND S. WATANABE (1981), *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam; Kodansha, Tokyo.

J. E. INGERSOLL, JR. (1987), *Theory of Financial Decision Making*, Rowman & Littlefield.

I. KARATZAS (1988), *On the pricing of American options*, Appl. Math. Optim., 17, pp. 37-60.

I. KARATZAS, P. LAKNER, J. P. LEHOCZKY, AND S. E. SHREVE (1988), *Dynamic equilibrium in a multi-agent economy: construction and uniqueness*, submitted.

I. KARATZAS, J. P. LEHOCZKY, S. P. SETHI, AND S. E. SHREVE (1986), *Explicit solution of a general consumption/investment problem*, Math. Oper. Res., 11, pp. 261-294.

I. KARATZAS, J. P. LEHOCZKY, AND S. E. SHREVE (1987), *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon*, SIAM J. Control Optim., 25, pp. 1557-1586.

——— (1988), *Existence and uniqueness of multi-agent equilibrium in a stochastic, dynamic consumption/investment model*, Math. Oper. Res., to appear.

I. KARATZAS AND S. E. SHREVE (1987), *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York.

P. LAKNER (1989), *Consumption/investment and equilibrium in the presence of several commodities*, Ph.D. dissertation, Columbia University, New York.

H. LATANÉ (1959), *Criteria for choice among risk ventures*, J. Political Economy, 67, pp. 144-155.

A. G. MALLIARIS AND W. A. BROCK (1982), *Stochastic Methods in Economics and Finance*, North-Holland, Amsterdam.

H. P. MCKEAN, JR. (1965), Appendix to P. A. Samuelson (1965): *A free boundary problem for the heat equation arising from a problem in mathematical economics*, Industr. Manag. Rev., 6, pp. 32-39.

R. C. MERTON (1969), *Lifetime portfolio selection under uncertainty: the continuous-time case*, Rev. Econom. Statist., 51, pp. 247-257.

——— (1971), *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3, pp. 373-413. Erratum, J. Ecom. Theory, 6 (1973), pp. 213-214.

——— (1973), *Theory of rational option pricing*, Bell J. Econom. Manag. Sci., 4, pp. 141-183.

S. R. PLISKA (1986), *A stochastic calculus model of continuous trading: optimal portfolio*, Math. Oper. Res., 11, pp. 371-382.

M. PONTIER AND M. PICQUÉ (1988), *Optimal portfolio for a small investor in a market model with discontinuous prices*, preprint.

P. A. SAMUELSON (1965), *Rational theory of warrant pricing*, Industr. Manag. Rev., 6, pp. 13-31.

——— (1969), *Lifetime portfolio selection by dynamic stochastic programming*, Rev. Econom. Statist., 51, pp. 239-246.

——— (1973), *Mathematics of speculative prices*, SIAM Rev., 15, pp. 1-42.

M. SCHWEIZER (1988), *Hedging of options in a general semimartingale model*, Ph.D. dissertation 8615, ETH, Zürich.

C. W. SMITH, JR. (1976), *Option pricing: a review*, J. Financial Econom., 3, pp. 3-51.

S. SUNDARESAN (1985), *Intertemporally dependent preferences in the theories of consumption, portfolio choice and equilibrium asset pricing*, preprint.

M. TAKSAR, M. J. KLASS, AND A. ASSAF (1988), *A diffusion model for optimal portfolio selection in the presence of brokerage fees*, Math. Oper. Res., 13, pp. 277-294.

M. TAQQU AND W. WILLINGER (1987), *The analysis of finite security markets using martingales*, Adv. Appl. Probab., 19, pp. 1-25.

E. O. THORP (1971), *Portfolio choice and the Kelly criterion*, In Stochastic Models in Finance, W. T. Ziemba and R. G. Vickson, eds., Academic Press, New York, pp. 599-619.

P. VAN MOERBEKE (1976), *On optimal stopping and free boundary problems*, Arch. Rational Mech. Anal., 60, pp. 101-148.

G. L. XU (1989), *A duality approach to the stochastic portfolio/consumption decision problem in a continuous-time market with short-selling restriction*, Ph.D. dissertation, Carnegie-Mellon University, Pittsburgh, PA.

T. ZARIPHOPOULOU (1989), *Optimal investment-consumption models with constraints*, Ph.D. dissertation, Brown University, Providence, RI.

REFERENCES ADDED IN PROOF

D. DUFFIE (1988), *Security Markets: Stochastic Models*, Academic Press, New York.

D. DUFFIE AND T. S. SUN (1986), *Transaction costs and portfolio choice in a discrete-continuous time setting*, preprint.

J. F. EASTHAM AND K. J. HASTINGS (1988), *Optimal impulse control of portfolios*, Math. Oper. Res., 13, pp. 588-605.

H. FÖLLMER AND D. SONDERMANN (1986), *Hedging of non-redundant contingent claims*, in Contributions to Mathematical Economics, W. Hildenbrand and A. Mas Collell, eds., pp. 205–223.

S. J. GROSSMAN AND J. L. VILA (1988), *Optimal dynamic hedging*, Princeton University, Princeton, NJ, preprint.

H. HE AND N. D. PEARSON (1988), *Consumption and portfolio policies with incomplete markets and short-sale constraints* I, *the finite-dimensional case*; II, *the infinite-dimensional case*, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, preprint.

I. KARATZAS, J. P. LEHOCZKY, S. E. SHREVE, AND G. L. XU (1989), *Utility maximization in an incomplete market*, in preparation.

H. E. LELAND (1985), *Option pricing and replication with transaction costs*, J. Finance, XL, pp. 1283–1301.

H. PAGÈS (1987), *Optimal consumption and portfolio policies when markets are incomplete*, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, preprint.

——— (1989), *Three essays in optimal consumption*, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1260–1278, November 1989

002

A SEQUENTIAL LINEAR PROGRAMMING ALGORITHM FOR SOLVING MONOTONE VARIATIONAL INEQUALITIES*

PATRICE MARCOTTE† AND JEAN-PIERRE DUSSAULT‡

Abstract. Applied to strongly monotone variational inequalities, Newton's algorithm achieves local quadratic convergence. In this paper it is shown how the basic Newton method can be modified to yield an algorithm whose global convergence can be guaranteed by monitoring the monotone decrease of the "gap function" associated with the variational inequality. Each iteration consists in the solution of a linear program in the space of primal-dual variables and of a linesearch. Convergence does not depend on strong monotonicity. However, under strong monotonicity and geometric stability assumptions, the set of active constraints at the solution is implicitly identified, and quadratic convergence is achieved.

Key words. mathematical programming, variational inequalities, nonlinear complementarity, Newton's method

AMS(MOS) subject classifications. 49D05, 49D10, 49D15, 49D35

0. Introduction. In this paper we consider the variational inequality problem defined on a convex compact polyhedron in R^n. Since this problem can be formulated as a fixed-point problem involving an upper semicontinuous mapping, it can be solved by simplicial or homotopy methods for which there already exists a vast literature (see Zangwill [16], Todd [14], Saigal [13]). For large-scale problems, however, these algorithms tend to become inefficient, both in terms of computer memory and running time requirements. This explains the renewed interest in algorithms closely related to procedures originally devised for iteratively solving systems of nonlinear equations (Ortega and Rheinboldt [10]) such as the Jacobi, Gauss–Seidel, and Newton schemes (see Pang and Chan [11], Josephy [5], Robinson [12]) or projection algorithms (Bertsekas and Gafni [2], Dafermos [4]) where the cost function is approximated, at each iteration, by a simpler, e.g., linear, separable, or symmetric function. Local or global convergence of the latter methods usually hinges on the a priori knowledge of lower bounds for the Lipschitz constant of the cost function, either in a neighborhood of a solution (for local convergence) or uniformly on the feasible domain (for global convergence). These conditions are difficult, while not impossible, to verify in practice.

Our approach is basically different. We choose as a merit function the complementary term (or *gap function*) associated with the primal-dual formulation of the variational inequality and find its global minimum by application of a first-order minimization algorithm. For monotone cost functions, we show that the algorithm converges globally to an equilibrium solution and possesses the finite termination property if the function is affine. Furthermore, under geometric stability and strong monotonicity assumptions, the algorithm implicitly identifies the set of constraints that are binding at the equilibrium solution, and convergence toward the equilibrium solution is quadratic. Numerical results comparing this method to Newton's method with and without linesearch (Marcotte and Dussault [9]) are provided.

* Received by the editors February 25, 1985; accepted for publication (in revised form) February 24, 1989.

† Département de Mathématiques, Collège Militaire Royal de Saint-Jean, Richelain, Québec, J0J 1R0, Canada. This research was supported by National Sciences and Engineering Research Council of Canada grants 5491 and 5789, and Academic Research Program of the Department of National Defense grant FUHBP.

‡ Département de Mathématiques et d'informatique, Université de Sherbrooke, Boul. Université, Sherbrooke, Québec, J1K 2R1, Canada.

1. Problem formulation. Notation and basic definitions. Let $\Phi = \{Bx \leqq b\}$, where B is an $m \times n$ matrix $(m > n)$, represent a nonempty convex compact polyhedron in R^n and let F be a continuously differentiable function from Φ into R^n with Jacobian F'. The *variational inequality problem* (VIP) associated with F and Φ consists in finding some vector x^* in Φ, called an *equilibrium solution*, satisfying the variational inequality (VI):

$$(x^* - x)^t F(x^*) \leqq 0 \tag{1}$$

for all x in Φ. Since an equilibrium solution is a fixed point of the upper semicontinuous mapping defined by $x \to T(x) = \{\arg\max_{y \in \Phi} (x-y)^t F(x)\}$ it follows from Kakutani's Theorem [6] and the compactness of Φ that the set S of equilibria is nonempty.

If the Jacobian $F'(x)$ is symmetric for all x in Φ then the function $F(x)$ is the gradient of some function $f: \Phi \to R^n$, and (1) is the mathematical expression of the first-order necessary conditions corresponding to the optimization problem:

$$\min_{x \in \Phi} f(x) = \int_0^x F(t)\, dt \tag{2}$$

where the line integral is independent of the path of integration and therefore unambiguously defined.

In order that a feasible point x be an equilibrium, it is necessary and sufficient that x be optimal for the linear program

$$\min_{y \in \Phi} y^t F(x). \tag{3}$$

The optimality conditions for (3) are met by x if and only if we have

$$\begin{aligned} &\lambda \geqq 0, \quad F(x) + B^t \lambda = 0 \quad \text{dual feasibility,} \\ &\lambda^t (Bx - b) = 0 \quad \text{complementary slackness,} \\ &Bx \leqq b \quad \text{primal feasibility.} \end{aligned} \tag{4}$$

In the following, (4) will be referred to as the *complementary formulation* of VIP. If F' is symmetric, (4) corresponds to the Kuhn–Tucker necessary optimality conditions for the optimization problem (2). If the constraint set Φ is not polyhedral, a formulation similar to (4) can be obtained by imposing a suitable constraint qualification condition on the problem. The constraints $Bx \leqq b$ will be referred to as the *structural constraints* associated with the variational inequality problem, and the constraints $F(x) + B^t \lambda = 0$, $\lambda \geqq 0$ as the *nonstructural constraints.*

DEFINITION 1. The function F is

(i) *Monotone* on Φ if $(x-y)^t (F(x) - F(y)) \geqq 0$ for all x, y in Φ;

(ii) *Strictly monotone* on Φ if $(x-y)^t (F(x) - F(y)) > 0$ for all x, y in Φ $(x \neq y)$;

(iii) *Strongly monotone* on Φ if there exists a positive number κ such that

$$(x-y)^t (F(x) - F(y)) \geqq \kappa \|x - y\|^2 \quad \text{for all } x, y \text{ in } \Phi.$$

When F is the gradient of some differentiable function f, then the various concepts of monotonicity previously defined correspond, respectively, to convexity, strict convexity, and strong convexity of f on Φ. For differentiable functions, we also have the following characterization (see Auslender [1]):

(i) Monotonicity on Φ: $(x-y)^t F'(x-y) \geqq 0$ for all x, y in Φ;

(ii) Strong monotonicity on Φ: $(x-y)^t F'(x)(x-y) \geqq \kappa \|x-y\|^2$ for all x, y in Φ, for some positive number κ.

The solution set S of (1) is nonempty, as noted earlier, convex if F is monotone, and a singleton if F is strictly monotone.

DEFINITION 2. The *gap function* associated with a VIP is defined, for x in Φ, as

$$g(x) = \max_{y \in \Phi} (x-y)^t F(x).$$

It is clear that a feasible point x is a solution of VIP if and only if it is a global minimizer for the gap function, i.e., $g(x) = 0$. Using this concept, VIP can be formulated as the linearly constrained optimization problem

$$\min_{x \in \Phi} g(x). \tag{5}$$

Although, in general, neither quasiconvex nor differentiable, it will be shown in Lemma 3 that any stationary point of (5) is an equilibrium solution. In particular, a globally convergent algorithm using the gap function as a merit function has been proposed by Marcotte [7].

DEFINITION 3. The *dual gap function* associated with VIP is defined as

$$\bar{g}(x) = \max_{y \in \Phi} (x-y)^t F(y).$$

The dual gap function is convex, but its evaluation requires the solution of a nonconvex (in contrast with linear for the gap function) mathematical program. Under a monotonicity assumption, any global minimizer of the problem $\min_{x \in \Phi} \bar{g}(x)$ is a solution to VIP. A solution algorithm based on direct minimization of the dual gap function can be found in Nguyen and Dupuis [17].

DEFINITION 4. We say that VIP is *geometrically stable* if $(y-x^*)^t F(x^*) \leqq 0$ for any equilibrium solution x^* implies that y lies in the optimal face T^*, i.e., the minimal face of Φ containing the set S of all solutions to VIP.

The above stability condition, especially useful when S is a singleton, ensures that T^* is stable under slight perturbations to the cost function F. It is implied by the generalization to VIP of the usual strict complementarity condition:

$$\{Bx^* = b \Leftrightarrow \lambda^* > 0\} \tag{6}$$

where λ^* is an optimal dual vector corresponding to x^* in the complementarity formulation (4). If F is strongly monotone, then geometric stability implies the strong regularity condition of Robinson [12]. Also, under geometric stability, there must exist at least one solution of VIP satisfying the strict complementarity condition (6); however it need not be unique, and there might exist optimal primal-dual couples that are not strictly complementary. Figure 1 provides examples where geometric stability holds while strict complementarity is not satisfied. In the first case, the problem is caused by a redundant constraint, while in the second case it is due to the linear dependence of the constraints' gradients at x^*.

2. Newton's algorithm. Since Newton's method is central to our local convergence analysis we recall its definition and main properties. Applied to VIP, Newton's method generates a sequence of iterates $\{x^k\}$ where x^0 is any vector in Φ and $x^{k+1} (k \geqq 0)$ is a solution to the VIP obtained by replacing F by its first-order Taylor expansion around x^k, i.e.,

$$(x^{k+1}-y)^t (F(x^k) + F'(x^k)(x-x^k)) \leqq 0 \quad \forall y \in \Phi. \tag{7}$$

The linearized problem will be denoted LVIP (x^k) and its (nonempty) set of solutions

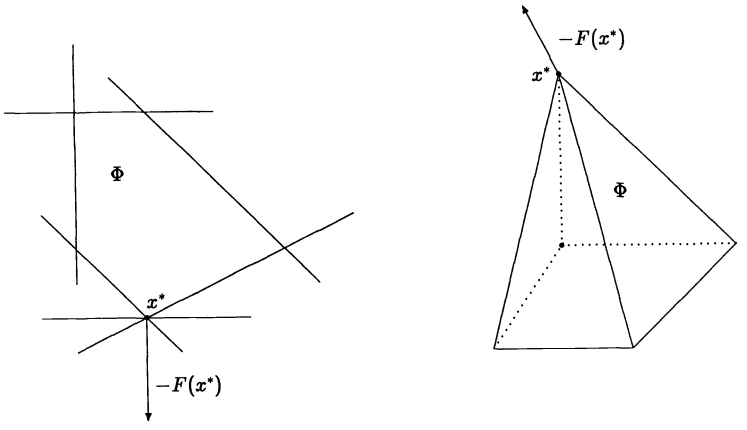

FIG. 1. *Geometric stability does not imply strict complementarity.* (a) *Redundant constraint.* (b) *Dependent constraints' gradients.*

NEW (x^k). The gap function associated with LVIP (x^k) (the *linearized gap function*) will be denoted $\mathrm{Lg}(x^k, x)$ and its mathematical expression is

$$\mathrm{Lg}(x^k, x) = \max_y (x-y)^t(F(x^k)+F'(x^k)(x-x^k)). \tag{8}$$

In a similar fashion we define the *linearized dual gap function* $\mathrm{L\bar{g}}(x^k, x)$:

$$\mathrm{L\bar{g}}(x^k, x) = \max_y (x-y)^t(F(x^k)+F'(x^k)(y-x^k)). \tag{9}$$

When F is strongly monotone and its F' is Lipschitzian, it can be shown that Newton's method is locally quadratically convergent. We quote Pang and Chan's [11] version of this result, also obtained by Josephy [5].

THEOREM 1. *If the matrix $F'(x^*)$ is positive definite and the function F' is Lipschitz continuous at x^* then there exists a neighborhood N of x^* such that if $x^k \in N$ then the sequence $\{x^k\}$ is well-defined and converges quadratically to x^*, i.e., there exists a constant ζ such that*

$$\|x^{k+1}-x^*\| \leqq \zeta \|x^k-x^*\|^2 \quad \forall k \text{ such that } x^k \in N \tag{10}$$

where $\|\cdot\|$ denotes the Euclidian norm in R^n.

The next result shows that Newton's algorithm has the capability of identifying T^*. Actually we will prove this result for a broad class of approximation algorithms where, at each iteration, x^{k+1} is defined as a solution to a VI where $F(x)$ is replaced by the function $G(x, x^k)$ parameterized in x^k and such that

(11) (i) $G(x, y)$ is strictly monotone in x;

(12) (ii) $G(x, y)$ is continuous as a function of (x, y);

(13) (iii) $G(x, x) = F(x)$.

Property (i) above ensures that x^{k+1} is unambiguously defined. Property (iii) ensures that if $x^{k+1} = x^k$ then x^k is the solution to the original VIP. In many practical situations, G is chosen as a strongly monotone function with symmetric Jacobian. Popular choices

for G are:

$G_i(x, y) = F_i(y_1, \cdots, y_{i-1}, x_i, y_{i+1}, \cdots, y_n), \quad i = 1, \cdots, n$ Jacobi iteration,

$G_i(x, y) = F_i(x_1, \cdots, x_i, y_{i+1}, \cdots, y_n), \quad i = 1, \cdots, n$ Gauss-Seidel iteration,

$G(x, y) = F(y) + F'(y)(x - y)$ Newton's method,

$G(x, y) = Ax + \rho[F(y) - Ay]$ Projection method,

where $\rho > 0$ and A is a symmetric positive definite matrix.

Other choices for G may be found in Pang and Chan [11] and Marcotte [8].

PROPOSITION 1. *Assume that F is monotone and that geometric stability holds for* VIP. *Let x^{k+1} be a solution to the* VI:

$$(x^{k+1} - y)^t G(x^{k+1}, x^k) \leqq 0 \quad \textit{for all } y \in \Phi$$

where G satisfies (11), (12), (13). *Then, for each optimal solution x^* of* VIP *there exists a neighborhood V of x^* such that if $x^k \in V$ then $x^{k+1} \in T^*$.*

Proof. Assume that the result does not hold. Then there exists an extreme point u of $\Phi - T^*$ and a subsequence $\{x^k\}_{k \in I}$ converging to some x^* such that

$$(x^{k+1} - u)^t G(x^{k+1}, x^k) = 0 \quad \text{for all } k \in I.$$

Taking the limit as $k \to \infty$ $(k \in I)$ we obtain

$$(x^* - u)^t F(x^*) = (x^* - u)^t G(x^*, x^*) \leqq 0$$

implying, by geometric stability, that $u \in T^*$, a contradiction. □

3. A linear approximation algorithm. In this section we present a model algorithm for solving VIP based on its complementarity formulation (4) that proceeds by successive linear approximations of both the objective and the nonstructural, usually nonlinear, constraints. Throughout this section the function F will be assumed monotone with Lipschitz continuous Jacobian F'.

Any solution to (4) is clearly a global minimizer for the following (usually) nonconvex, nonlinearly constrained mathematical program:

$$\begin{aligned} &\min_{x,\lambda} h(x, \lambda) \overset{\text{def}}{=} \lambda^t(b - Bx) = x^t F(x) + b^t \lambda \\ &\text{subject to} \quad F(x) + B^t \lambda = 0, \quad Bx \leqq b, \quad \lambda \geqq 0. \end{aligned} \tag{14}$$

The following lemma relates the objective in (14) to the gap function.

LEMMA 1. *We have*

$$\begin{aligned} &g(x) = \min_{\lambda} h(x, \lambda) \\ &\textit{subject to} \quad F(x) + B^t \lambda = 0, \quad \lambda \geqq 0. \end{aligned} \tag{15}$$

Proof.

$$\begin{aligned} g(x) &= \max_{y \in \Phi} (x - y)^t F(x) \\ &= x^t F(x) - \min_{By \leqq b} y^t F(x) \\ &= x^t F(x) - \max_{\substack{B^t \mu = F(x) \\ \mu \leqq 0}} b^t \mu \end{aligned}$$

by linear programming duality theory. Hence

$$g(x) = x^t F(x) + \min_{\substack{F(x)+B^t\lambda=0 \\ \lambda \geqq 0}} b^t\lambda$$

after setting $\lambda = -\mu$, and the result follows if we replace $F(x)$ by the equivalent term $-B^t\lambda$. □

The next lemma, basic to our global convergence analysis, states that any stationary point of the mathematical program (14) is actually an equilibrium solution to VIP and justifies the use of an algorithm based on identifying points satisfying first-order conditions of (14). The proof does not rely on any sort of constraint qualification for the nonlinearly constrained problem (14).

LEMMA 2. *Let* $(\bar{x}, \bar{\lambda})$ *be a vector satisfying the first-order necessary optimality conditions for* (14). *Then* $\bar{x}$ *is a solution to* VIP.

Proof. It suffices to show that $h(\bar{x}, \bar{\lambda}) = 0$. Assume that $h(\bar{x}, \bar{\lambda}) > 0$. Without loss of generality we also assume that $h(\bar{x}, \bar{\lambda}) = g(\bar{x})$; otherwise, $\bar{\lambda}$ would not be optimal for the linear program

$$\begin{aligned} &\min_{\lambda} \bar{h}(x, \lambda) \\ &\text{subject to} \quad F(\bar{x}) + B^t\lambda = 0, \quad \lambda \geqq 0 \end{aligned} \tag{16}$$

and an optimal λ-solution to (16) would constitute, together with $\bar{x}$, an obvious descent direction for h at $(\bar{x}, \bar{\lambda})$.

Consider the linearized problem LVIP $(\bar{x})$ with its gap function Lg $(\bar{x})$ and complementarity formulation:

$$\begin{aligned} &\min_{x,\lambda} \bar{h}(x, \lambda) \overset{\text{def}}{=} x^t[F(\bar{x}) + F'(\bar{x})(x - \bar{x})] + b^t\lambda \\ &\text{subject to} \quad F(\bar{x}) + F'(\bar{x})(x - \bar{x}) + B^t\lambda = 0, \quad \lambda \geqq 0. \end{aligned} \tag{17}$$

Problem (17) constitutes a positive semidefinite quadratic program whose optimal solution's primal vector corresponds to a (not necessarily unique) Newton direction. Consider a Frank-Wolfe direction $d = (\tilde{x} - \bar{x}, \tilde{\lambda} - \bar{\lambda})$ for (17) at the point $(\bar{x}, \bar{\lambda})$. Direction d is a feasible descent direction for the linearized gap function Lg at $\bar{x}$. Since $\nabla h(\bar{x}, \bar{\lambda})$ is identical with $\nabla \bar{h}(\bar{x}, \bar{\lambda})$, and so are the directional derivatives of Lg and g, it follows that $\tilde{x} - \bar{x}$ is also a feasible descent direction for g at $\bar{x}$. □

We are now in a position to give a precise statement of our algorithm.

ALGORITHM N^*.
Initialization.
Let x^0 be any vector in Φ and

$$\lambda^0 \in \arg\min_{\lambda \geqq 0} b^t\lambda$$

$$\text{subject to} \quad F(x) + B^t\lambda = 0$$

and set $k \leftarrow 1$.
while convergence criterion not met
do 1) Find descent direction *d*.

Let $(d_x(x^k), d_\lambda(x^k))$ be an *extremal* solution to the linear problem

$$\min_{\substack{x\in\Phi\\ \lambda\geqq 0}} x^t(F(x^k)+F'^t(x^k)x^k)+b^t\lambda$$

$$\text{subject to} \quad F(x^k)+F'(x^k)(x-x^k)+B^t\lambda=0. \tag{18}$$

Set $d\leftarrow(d_x(x^k)-x^k, d_\lambda(x^k)-\lambda_k)$.

2) Perform arc search on the gap function.

$$\text{if } g(d_x(x^k))\leqq\tfrac{1}{2}g(x^k) \text{ then } \bar\theta\leftarrow 1$$
$$\text{else } \bar\theta\in\arg\min_{\theta\in[0,1]} g[x^k+\theta(d_x(x^k)-x^k)]. \tag{19}$$

3) Update.

$$x^{k+1}\leftarrow x^k+\bar\theta(d_x(x^k)-x^k)$$

$$\lambda^{k+1}\in\arg\min_{\substack{\lambda\geqq 0\\ F(x^{k+1})+B^t\lambda=0}} b^t\lambda$$

$$k\leftarrow k+1$$

endwhile.

Some comments are in order:

(1) At step 2) of Algorithm N^*, the minimization, with respect to the primal vector x, of the nondifferentiable objective g could be seen as a search along an arc in the space of primal-dual variables (x, λ). Since dual vectors λ have to be computed repeatedly, this operation can be carried out efficiently using reoptimization techniques of linear programming.

(2) It is not required, or even advisable, that the arc search be carried out exactly. For instance, the Armijo–Goldstein stepsize rule, or any rule guaranteeing a "sufficient" decrease of the objective along the search direction could be implemented.

(3) For affine functions $F=Ax+a$, Algorithm N^* reduces to the standard Frank–Wolfe procedure for solving quadratic programming problems, as then the nonstructural constraints become linear.

4. Convergence analysis. We first state and prove a global convergence result for Algorithm N^*.

PROPOSITION 2. *Any point of accumulation of a sequence generated by Algorithm N^* is an equilibrium solution.*

Proof. If $g(d_x(x^k))\leqq .5g(x^k)$ infinitely often at (19), then $\lim_{k\to\infty} g(x^k)=0$. Otherwise the linesearch in (19) is asymptotically always performed and, to prove global convergence, we will strive to check the conditions behind Zangwill's global convergence theorem, namely:

(i) All points generated by the algorithm lie in a compact set.

(ii) The algorithmic map is closed outside the set of solution points S.

(iii) At each iteration, strict decrease of the objective function occurs.

(i) Since Φ is compact by assumption, it is sufficient to show that the sequence $\{d_\lambda(x^k)\}$ is bounded. By definition of the sequence $\{d_\lambda(x^k)\}$ we have

$$d_\lambda(x^k)\in\arg\min_{\lambda\geqq 0} b^t\lambda$$

$$\text{subject to} \quad B^t\lambda=H(x^k) \tag{20}$$

where $H(x^k)\overset{\text{def}}{=}F(x^k)+F'(x^k)(d_x(x^k)-x^k)\in R^n$.

First observe that the linear program:

$$\min_{\lambda \geqq 0} b^t\lambda \qquad \text{subject to} \quad B^t\lambda = -H(x^k) \tag{21}$$

is the dual of the linear program

$$\max_{x \in \Phi} -x^t F(x^k) \tag{22}$$

that is feasible and bounded; hence, by linear programming duality, we have that (17) is also feasible and bounded, i.e., that (17) possesses at least one optimal basic solution. Let $\{N_e\}_{e=1,\cdots,p}$ denote the set of full rank square submatrices (basis) of B^t. Since $d_\lambda(x^k)$ is extremal, we have

$$d_\lambda(x^k) = -N_e^{-1}H(x^k)$$

for some $e \in \{1, \cdots, p\}$. From the continuity of F and F' we deduce that $H(x^k)$ must lie in some compact set K independent of x^k. Therefore $d_\lambda(x^k) \in C \stackrel{\text{def}}{=} \cup_{e=1}^p -N_e^{-1}K$, which is bounded. The same continuity argument is then used to show boundedness of the sequence $\{\lambda^{k+1}\}$.

(ii) The closedness of the algorithmic map follows directly from the continuity of F' and the closedness of the linesearch strategy used.

(iii) We must prove that $h(x^{k+1}, \lambda^{k+1}) < h(x^k, \lambda^k)$ if the latter term is positive (not zero). This is a direct consequence of Lemma 2. □

PROPOSITION 3. *If F is monotone on Φ and affine, then Algorithm $N^\#$ converges in a finite number of iterations.*

Proof. Replacing F by $Ax + a$ in (16) yields a quadratic programming problem. Its solution set is a face $\bar{T}$ of the polyhedron $\{Ax + B^t\lambda = 0,\ Bx \leqq b,\ \lambda \geqq 0\}$. For some iterate k we must have that $(d_x(x^k), d_\lambda(x^k))$ lies in $\bar{T}$ (otherwise the iterates would always be bounded away from $\bar{T}$, contradicting global convergence of the method). When $(d_x(x^k), d_\lambda(x^k)) \in \bar{T}$ we have $\bar{\theta} = 1$ and $(x^{k+1}, \lambda^{k+1}) \in \bar{T}$. □

Remark. The preceding result is also valid under the assumption that T^* is a singleton (F monotone but not necessarily affine). The proof is similar.

To obtain a rate-of-convergence result for Algorithm $N^\#$ we assume, until explicitly stated otherwise, that the function F is strongly monotone in a neighborhood of the solution x^* with strong monotonicity coefficient κ and that the geometric stability condition is satisfied at x^*. This implies that the entire sequence $\{x^k\}$ converges to the unique solution x^*. Under these assumptions we will show that Algorithm $N^\#$ is locally equivalent to Newton's method, thus implying quadratic convergence and implicit identification of the set of active constraints at x^*. We first show that the descent direction d obtained from Algorithm $N^\#$ satisfies $d_x(x^k) = \text{NEW}(x^k)$ if x^k is sufficiently close to x^*. The following lemmas will be used in the proof.

LEMMA 3. *The optimal dual vector $y(x^k)$ associated with the nonstructural constraint $F(x) + B^t\lambda = 0$ of (18) satisfies $\lim_{k\to\infty} y(x^k) = x^*$.*

Proof. Write the Lagrangian dual of the linear program (18):

$$\max_{y} \min_{\substack{x \in \Phi \\ \lambda \geqq 0}} x^t[F(x^k) + F'^t(x^k)x^k] + b^t\lambda - y^t[F(x^k) + F'(x^k)(x - x^k) + B^t\lambda].$$

Then observe that the inner minimum has value $-\infty$ unless $By \leqq b$, in which case the minimum over nonnegative λ is achieved when λ is zero, yielding

$$\max_{y \in \Phi} \min_{x \in \Phi} x^t[F(x^k) + F'^t(x^k)x^k] - y^t[F(x^k) + F'(x^k)(x - x^k)].$$

This expression is equivalent, modulo a constant term, to

$$\max_{y\in\Phi}\min_{x\in\Phi}(x-y)^t[F(x^k)+F'(x^k)(x-x^k)]-(x-x^k)F'(x^k)(x-x^k) \tag{23}$$

and constitutes a quadratic perturbation of the dual gap function $L\bar{g}$ at x^k. Since $y(x^k)$ is dual-optimal for (18) it must correspond to the y-part of a solution to (23). If $y(x^k)$ does not converge to x^* then there exists a subsequence $\{x^k\}_{k\in I}$ such that $\lim_{k\to\infty,k\in I} y(x^k)=\tilde{y}\neq x^*$. Passing to the limit in (23) we obtain, after setting x to $\tilde{x}$:

$$\begin{aligned}
&\lim_{\substack{k\to\infty\\ k\in I}}\max_{y\in\Phi}\min_{x\in\Phi}(x-y)^t[F(x^k)+F'(x^k)(x-x^k)]-(x-x^k)^tF'(x^k)(x-x^k)\\
&\quad\leqq(\tilde{y}-\tilde{y})^t[F(x^*)+F'(x^*)(\tilde{y}-x^*)]-(\tilde{y}-x^*)^tF'(x^*)(\tilde{y}-x^*)\\
&\quad\leqq-\kappa\|\tilde{y}-x^*\|^2 \quad\text{by strong monotonicity}\\
&\quad<0.
\end{aligned}$$

But this contradicts the optimality of the sequence $\{y(x^k)\}_{k\in I}$ since we obtain, by taking $y=x^*$:

$$\begin{aligned}
&\lim_{k\to\infty}\min_{x\in\Phi}(x-x^*)^t[F(x^k)+F'(x^k)(x-x^k)]-(x-x^k)^tF'(x^k)(x-x^k)\\
&\quad=\min_{x\in\Phi}(x-x^*)^tF(x^*)\\
&\quad=0 \quad\text{by definition of } x^*.
\end{aligned}$$

□

LEMMA 4. *There exists an index K such that $k\geqq K$ implies $d_x(x^k)\in T^*$.*

Proof. From (23) we get

$$\begin{aligned}
d_x(x^k)\in D(x^k)\overset{\text{def}}{=}\arg\min_{x\in\Phi}\,&(x-y(x^k))^t[F(x^k)+F'(x^k)(x-x^k)]\\
&-(x-x^k)^tF'(x^k)(x-x^k).
\end{aligned} \tag{24}$$

Since $y(x^k)\to x^*$ as $k\to\infty$ (Lemma 3), (24) represents, for x^k close to x^*, a small quadratic perturbation of the linear program: $\min_{x\in\Phi}x^tF(x^*)$. It follows from the geometric stability assumption that $d_x(x^k)\in T^*$. □

COROLLARY. $d_x(x^*)\in T^*$.

Proof. Since F' is continuous, the point-to-set mapping $\bar{x}\to\{d_x(\bar{x})\}$ is upper semicontinuous. Hence $d_x(x^*)\in\{d_x(\lim_{k\to\infty}x^k)\}=\lim_{k\to\infty}\{d_x(x^k)\}\in T^*$. □

LEMMA 5. $\lim_{k\to\infty}d_x(x^k)=x^*$.

Proof. From the proof of Lemma 2, we have

$$g'(x^k;\,d_x(x^k)-x^k)<0.$$

Passing to the limit and using upper semicontinuity there comes

$$g'(x^*;\,d_x(x^*)-x^*)\leqq 0.$$

But, by Danskin's rule of differentiation of max-functions (see [18]), we have

$$g'(x^*;\,d_x(x^*)-x^*)=\max_{y\in T^*}[d_x(x^*)-x^*]^t[F(x^*)-F'^t(x^*)(y-x^*)].$$

Assume that T^* is not the singleton x^* (otherwise the result follows trivially from Lemma 5) and let ε be a positive number such that $\hat{y}\overset{\text{def}}{=}x^*-\varepsilon(d_x(x^*)-x^*)\in T^*$ (see

Fig. 2). Then we have

$$\begin{aligned} 0 &\geqq g'(x^*;\, d_x(x^*) - x^*) \\ &\geqq \varepsilon[d_x(x^*) - x^*]'[F(x^*) + F''(x^*)(d_x(x^*) - x^*)] \\ &\geqq \varepsilon\kappa \|d_x(x^*) - x^*\|^2, \end{aligned}$$

implying that $d_x(x^*) = x^*$. □

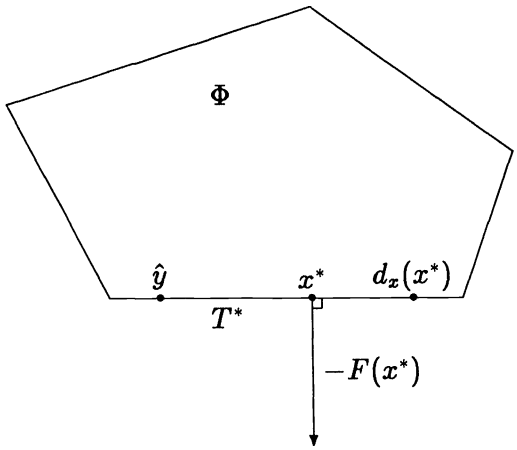

FIG. 2

LEMMA 6. *There exists an index K such that for* $k \geqq K$, $d_x(x^k) = \text{NEW}(x^k)$.

Proof. From (18), $d_\lambda(x^k)$ is an optimal dual vector for the linear program

$$\min_{z \in \Phi} z'[F(x^k) + F'(x^k)(d_x(x^k) - x^k)]. \tag{25}$$

For k large, $d_x(x^k)$ is close to x^* (Lemma 5) and problem (25) is an arbitrary small perturbation of the linear program

$$\min_{z \in \Phi} z'F(x^*)$$

whose set of optimal solutions is T^*, by definition. Therefore the optimal solutions to (25) lie in T^* by geometric stability. From the complementary slackness theorem of linear programming we can write

$$d_\lambda(x^k)'(Bd_x(x^k) - b) = 0.$$

We conclude that the couple $(d_x(x^k), d_\lambda(x^k))$ is optimal for the quadratic program (17). Since its solution is unique in x and equal by definition to NEW (x^k) we conclude that $d_x(x^k) = \text{NEW}(x^k)$. □

PROPOSITION 4. *There exist positive constants* α *and* β *such that*

$$\alpha \|x - x^*\| \leqq g(x) \leqq \beta \|x - x^*\|.$$

Proof. It suffices to prove the result in a neighborhood of x^*.

(26) (i) Proof that $g(x) \leqq \beta \|x - x^*\|$.

$$\begin{aligned} g(x) &= \max_{y \in \Phi} (x-y)^t F(x) \\ &= (x-x^*)^t F(x) + \max_{y \in \Phi} (x^*-y)^t F(x) \\ &\leqq \|x-x^*\| \cdot \|F(x)\| + \max_{y \in \Phi} (x^*-y)^t F(x^*) \\ &\quad + \max_{y \in \Phi} (x^*-y)^t (F(x)-F(x^*)) \\ &\leqq \|x-x^*\| \cdot \|F(x)\| + D\|x-x^*\| \sup_{\xi \in \Phi} \|F'(\xi)\| \\ &\leqq (M + M' \operatorname{diam}(\Phi)) \|x-x^*\| \end{aligned}$$

where $M \overset{\text{def}}{=} \sup_{x \in \Phi} \|F(x)\|$, $M' \overset{\text{def}}{=} \sup_{\xi, \eta \in \Phi, \xi \neq \eta} \|F(\xi) - F(\eta)\| / \|\xi - \eta\|$, and D is the diameter of Φ. Then set $\beta = M + M'D$.

(27) (ii) Proof that $g(x) \geqq \alpha \|x - x^*\|$.

We consider three mutually exclusive cases.

Case 1. $T^* = x^*$. (Fig. 3.) For x sufficiently close to x^* we have

$$\begin{aligned} g(x) &= (x-x^*)^t F(x) \\ &= (x-x^*)^t F(x^*) + (x-x^*)^t (F(x) - F(x^*)) \\ &\geqq \|x-x^*\| \cdot \|F(x^*)\| \cos(x-x^*, F(x^*)) + \kappa \|x-x^*\|^2. \end{aligned}$$

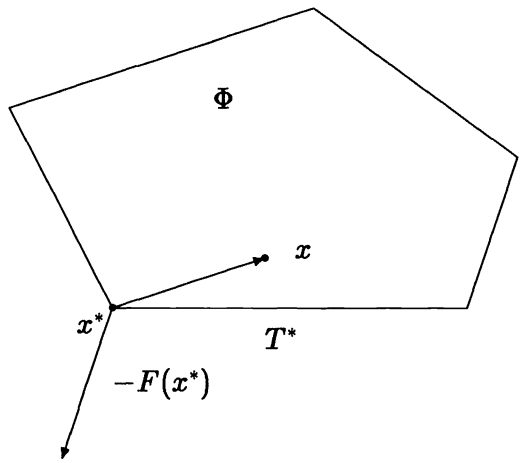

FIG. 3. *Case* 1.

Since $F(x^*)$ is orthogonal to no feasible direction from x^* into Φ (by geometric stability) we must have that $\cos(x-x^*, F(x^*))$ is positive and bounded away from zero. Hence (27) holds with

$$\alpha = \bar{\alpha} \overset{\text{def}}{=} \inf_{\substack{\xi \in \Phi \\ \xi \neq x^*}} \{\cos(\xi - x^*, F(x^*))\} \|F(x^*)\| > 0.$$

Case 2. $T^* \neq \{x^*\}$ and $x \in T^*$. (Fig. 4.) Let ρ be a positive number such that the mapping $\text{Proj}_\Phi(x-(1/\rho)F(x))$, defined for $x \in \Phi$, is contracting, where Proj_Φ denotes the projection operator on Φ, in the usual Euclidean norm. The existence of such a number ρ is a consequence of, say, Example 3.1 of Dafermos [4]. For x sufficiently

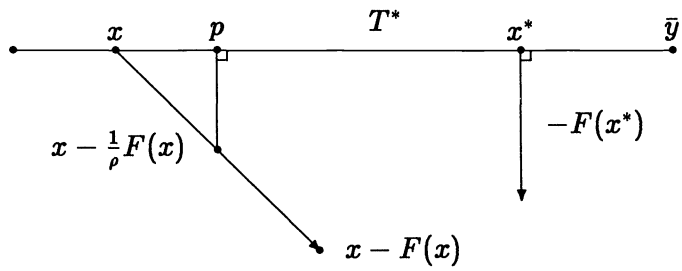

FIG. 4. *Case* 2.

close to x^*, $p \overset{\text{def}}{=} \text{Proj}_\Phi(x-(1/\rho)F(x))$ lies in T^* (see Proposition 1). Let $\theta \in [0,1]$ be the contraction constant, dependent on ρ; we have $\|p-x^*\| \leqq \theta\|x-x^*\|$. We have

$$\begin{aligned}\|p-x\| &\geqq \|x-x^*\| - \|x^*-p\| \quad \text{by the triangle inequality} \\ &\geqq (1-\theta)\|x-x^*\|.\end{aligned} \tag{28}$$

Also by construction of p

$$(x-p)^t F(x) = \rho\|x-p\|^2. \tag{29}$$

Define

$$\psi = \max\{\phi \mid x+\phi(p-x) \in T^*\} \tag{30}$$

(ψ must be positive since x lies in the relative interior ri (T^*)) and $\bar{y} = x+\psi(p-x)$. We have

$$\begin{aligned}(x-\bar{y})^t F(x) &= \psi(x-p)^t F(x) \\ &= \rho\psi\|x-p\|^2 \quad \text{by (29)} \\ &\geqq \rho\psi\|x-p\|(1-\theta)\|x-x^*\| \quad \text{by (28)}.\end{aligned}$$

Now $\psi\|x-p\| = \|x-\bar{y}\|$ must be bounded from below by some positive number s since x lies in ri (T^*) and $\bar{y}$ is on the boundary of T^*. It follows that

$$\begin{aligned}g(x) &\geqq (x-\bar{y})^t F(x) \\ &\geqq \rho(1-\theta)s\|x-x^*\|\end{aligned}$$

and the result holds with $\alpha = \rho s(1-\theta)$.

Case 3. $x \notin T^*$ (consequently $T^* \neq \Phi$). (Fig. 5.) Define $p = \text{Proj}_{T^*}(x)$. First we will show that $\cos(x-p, F(x^*))$ is bounded below by some positive number γ.

Define, for $x \notin T^*$, the function $x \to \eta(x)$ where $\eta(x)$ is the intersection of the line going through the segment $[p, x]$ with the boundary of Φ, in the direction $x-p$. Let $B_\varepsilon(x^*)$ be a ball of radius ε about x^*, $H = B_\varepsilon(x^*) \cap \Phi - T^*$, and E the closure of $\eta(H)$ (see Fig. 6). We have $E \cap T^* = \varnothing$ and

$$\begin{aligned}\cos(x-p, F(x^*)) &= \cos(\eta(x)-p, F(x^*)) \\ &\geqq \min_{v \in E} \cos(v-p, F(x^*)).\end{aligned} \tag{31}$$

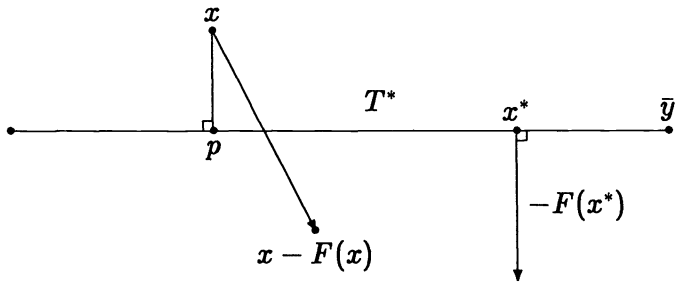

FIG. 5. *Case* 3.

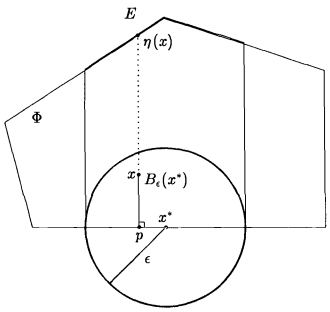

FIG. 6

But $\cos(v-p, F(x^*))>0$ $(p\in T^*)$ for each $v\notin T^*$ by geometric stability. Hence, $\cos(x-p, F(x^*))\geqq\gamma>0$.

We then write $(x-p)^tF(x^*)\geqq\gamma\|x-p\|\cdot\|F(x^*)\|$. Thus

$$(x-p)^tF(x)\geqq\frac{\gamma}{2}\|x-p\|\cdot\|F(x^*)\| \tag{32}$$

for x sufficiently close to x^*. Now consider the following two subcases.

Case 3.1. $\|x-p\|\leqq\zeta\|p-x^*\|$ with $\zeta=\rho s(1-\theta)/2\theta DM'$. Define $\bar{y}$ as in Case 2 (see Fig. 4). Then

$$\begin{aligned}
g(x)&\geqq(x-\bar{y})^tF(x)\\
&=(x-p)^tF(x)+(p-\bar{y})^tF(x)\\
&=0+(p-\bar{y})^tF(p)+(p-\bar{y})^t(F(x)-F(p))\\
&\geqq\rho s(1-\theta)\|x-x^*\|-\zeta\|p-x^*\|DM' \quad \text{since } p\in T^* \quad \text{(see Case 2)}\\
&\geqq(\rho s(1-\theta)-\zeta\theta DM')\|x-x^*\|\\
&\geqq\frac{\rho s(1-\theta)}{2}\|x-x^*\|.
\end{aligned}$$

Set $\alpha=\rho s(1-\theta)/2$.

Case 3.2. $\|x-p\|\geqq\zeta\|p-x^*\|$. We have

$$\|x-x^*\|\leqq\|x-p\|+\|p-x^*\|\leqq(1+\zeta)\|x-p\|. \tag{33}$$

We obtain

$$g(x) \geqq (x-p)^t F(x)$$

$$\geqq \frac{\gamma}{2}\|x-p\| \cdot F(x^*) \quad \text{by (32)}$$

$$\geqq \frac{\gamma}{2}\frac{1}{1+\zeta}\|x-x^*\| \cdot \|F(x^*)\|$$

by (28) and (29), with $F(x^*) \neq 0$, and the result holds with $\alpha = \gamma\|F(x^*)\|/2(1+\zeta)$. □

Remark 1. The above general proof does not require differentiability of the cost mapping F. If F is differentiable, the proof of Proposition 4 can be somewhat streamlined (see Dussault and Marcotte [21]).

Remark 2. Proposition 4 strengthens a result of Pang [19] who derives an estimate of the form

$$\|x-x^*\| \leqq \omega\sqrt{g(x)}$$

for some positive constant ω.

PROPOSITION 5. *Let $\{x^k\}$ be a sequence generated by Algorithm N^*. Then there exists an index K such that for $k \geqq K$, $x^{k+1} = \text{NEW}(x^k)$, the Newton iterate.*

Proof. We must prove that $g(\text{NEW}(x^k)) \leqq \frac{1}{2}g(x^k)$ for $k \geqq K$, in which case Algorithm N^* will set x^{k+1} to $d_x(x^k)$, which is equal to $\text{NEW}(x^k)$ by Lemma 6:

$$g(\text{NEW}(x^k)) \leqq \beta\|\text{NEW}(x^k)-x^*\| \quad \text{by Proposition 4}$$

$$\leqq \beta c\|x^k-x^*\|^2 \quad \text{from (10)}$$

$$\leqq \frac{\beta c}{\alpha}\|x^k-x^*\|g(x^k) \quad \text{by Proposition 4}$$

$$\leqq \frac{1}{2}g(x^k)$$

as soon as $\|x^k-x^*\| \leqq \alpha/2\beta c$. □

The preceding results can be summarized in a theorem.

THEOREM 2. *Consider a* VIP *with monotone cost function F and let $\{x^k\}$ be a sequence generated by Algorithm N^*. If* VIP *is geometrically stable, then*

(i) $g(x^{k+1}) < g(x^k)$ *if* $g(x^k) \neq 0$.

(ii) $\lim_{k\to\infty} g(x^k) = 0$.

(iii) *If F is affine or T^* is a singleton then there exists an index K such that $g(x^k) = 0$ for $k \geqq K$ (finite convergence).*

(iv) *If F is strongly monotone then the sequence $\{x^k\}$ converges quadratically to the point x^* and there exists an index K such that $x^k \in T^*$ whenever $k \geqq K$.*

5. Numerical results. A working version of Algorithm N^* has been developed, using a standard linear programming code, and contrasted against Newton's method, with or without linesearch. The asymmetric linear complementarity subproblems in Newton's method have been solved by Lemke's Complementary Pivoting Algorithm.

In the test problems, Φ has been taken as the unit simplex $\sum_{i=1}^{n} x_i = 1$, $x \geqq 0$ and the mapping F assumed the general form

$$F(x) = (A - A^t)x + B^t Bx + \gamma C(x) + b$$

where the entries of matrices A and B are randomly generated uniform variates, $C(x)$ is a nonlinear diagonal mapping with components $C_i(x) = \arctan(x_i)$, and the constant vector b is chosen such that the exact optimum be known a priori. The parameter γ is used to vary the asymmetry and nonlinearity of the cost function.

Sixteen five-dimensional and sixteen 15-dimensional problems have been generated, with γ-values ranging from 10-40. Newton's search direction differs from Algorithm $N^{\#}$'s direction in 18 of the 32 problems. In some instances (Figs. 7-10) Algorithm $N^{\#}$ yields a direction as good or better than Newton's direction. For some other problems (Figs. 11-14) Newton's direction is slightly superior. In all cases, the difference in the number of iterations required to achieve a very low gap value is small.

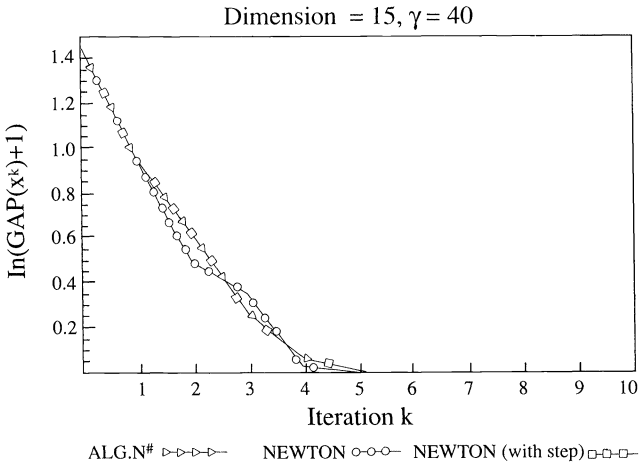

FIG. 7

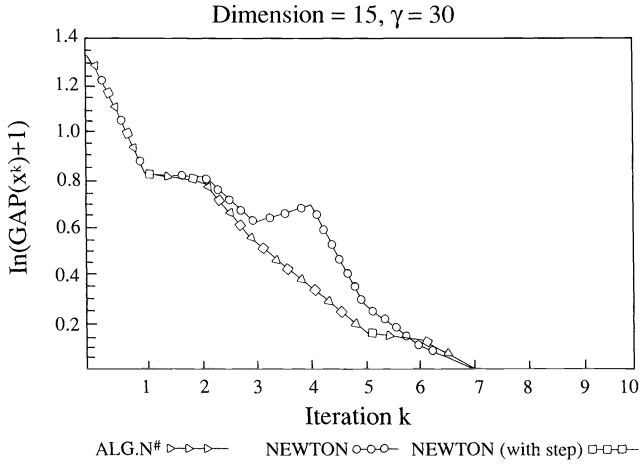

FIG. 8

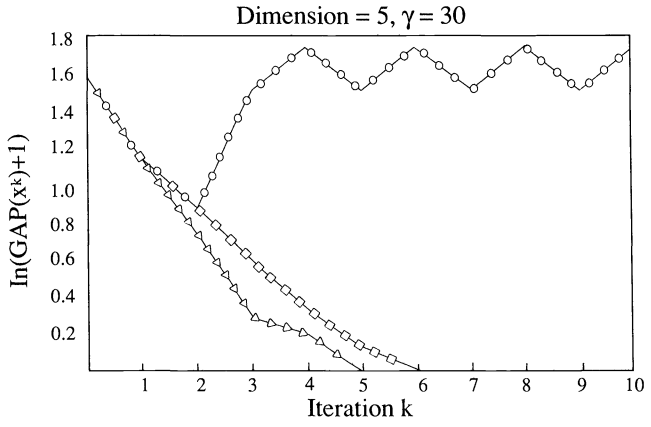

FIG. 9

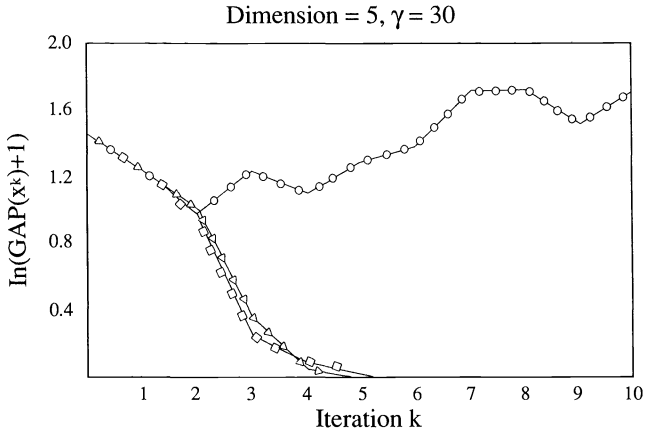

FIG. 10

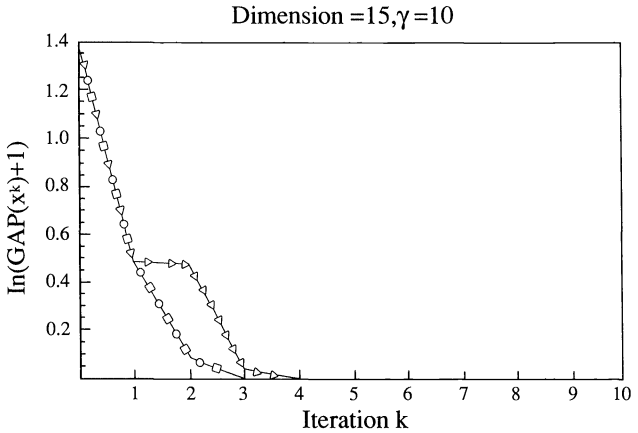

FIG. 11

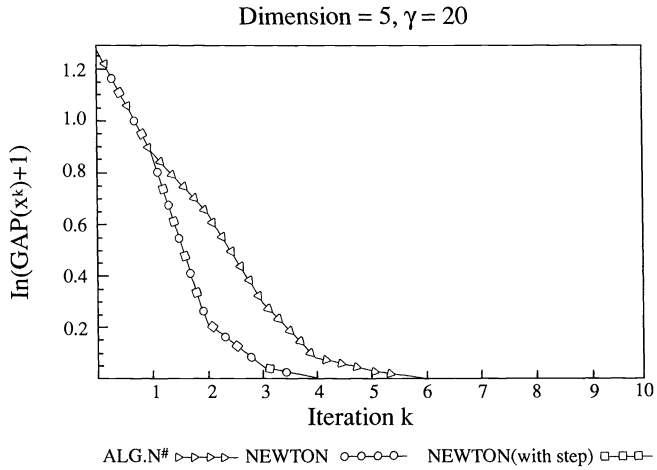

FIG. 12

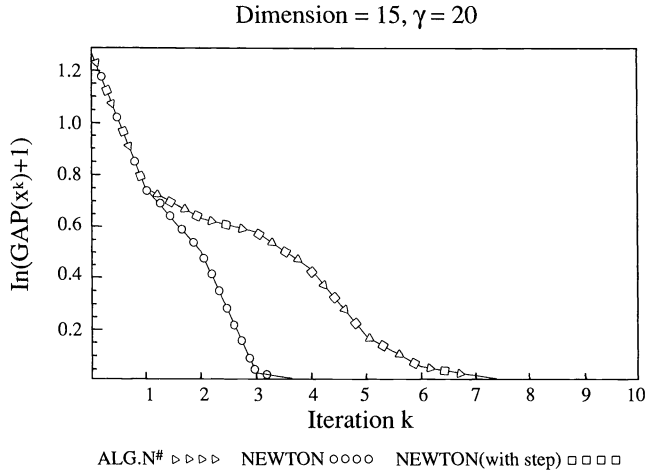

FIG. 13

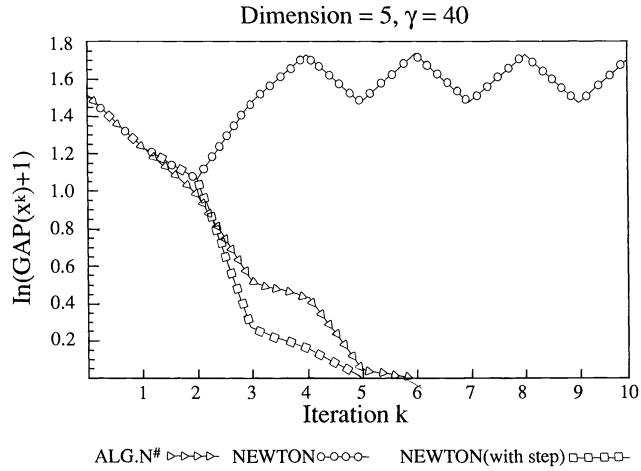

FIG. 14

This preliminary testing shows some promise for the linearization algorithm. Its direction finding subproblem involves a linear program, versus an asymmetric linear complementarity problem for Newton's method. The linear subproblem bears close resemblance to the linear program that must be solved to evaluate the gap function, and as such could benefit from some fine tuning of the computer code. Moreover, it may well prove unnecessary to solve the subproblem exactly, yielding another area for further improvement. In contrast, solving linear complementarity problems yields a feasible solution only at termination, therefore making the implementation of an inexact strategy more difficult.

Finally let us mention that Marcotte and Guélat [20] have successfully implemented Algorithm N^* to solve large-scale network equilibrium problems when the mapping F, i.e., its Jacobian matrix, is highly asymmetric.

6. Conclusion. The main result of this paper has been to prove global and quadratic convergence of an algorithm for solving monotone variational inequalities. The algorithm operates by solving linear programs in the space of primal-dual variables. Computational experiments show that the algorithm is efficient for solving both small-scale and large-scale problems.

Acknowledgments. The authors are indebted to anonymous referees for relevant comments on an earlier version of this paper that led to numerous improvements.

REFERENCES

[1] A. AUSLENDER, *Optimisation. Méthodes numériques*, Masson, Paris, 1976.
[2] D. BERTSEKAS AND E. M. GAFNI, *Projection methods for variational inequalities with application to the traffic assignment problem*, Math. Programming Stud., 17 (1982), pp. 139–159.
[3] R. W. COTTLE AND G. B. DANTZIG, *Positive (semi) definite programming*, in Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1967.
[4] S. C. DAFERMOS, *An iterative scheme for variational inequalities*, Math. Programming, 26 (1983), pp. 40–47.
[5] N. H. JOSEPHY, *Newton's method for generalized equations*, Technical Report 1966, Mathematical Research Center, University of Wisconsin, Madison, WI, 1979.
[6] S. KAKUTANI, *A generalization of Brouwer's fixed point theorem*, Duke Math. J., 8 (1941), pp. 457–459.
[7] P. MARCOTTE, *A new algorithm for solving variational inequalities, with application to the traffic assignment problem*, Math. Programming, 33 (1985), pp. 339–351.
[8] ———, *Algorithms for the network oligopoly problem*, J. Oper. Res. Soc., 38 (1987), pp. 1051–1065.
[9] P. MARCOTTE AND J.-P. DUSSAULT, *A note on a globally convergent method for solving monotone variational inequalities*, Oper. Res. Lett., 6 (1987), pp. 35–42.
[10] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[11] J. S. PANG AND D. CHAN, *Iterative methods for variational and complementarity problems*, Math. Programming, 24 (1982), pp. 284–313.
[12] S. M. ROBINSON, *Generalized equations*, in Mathematical Programming: The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, New York, 1983, pp. 346–367.
[13] R. SAIGAL, *Fixed point computing methods*, in Operations Research Support Methodology, A. G. Holzman, ed., Marcel Dekker, New York, 1979.
[14] M. J. TODD, *The Computation of Fixed Point and Applications*, Springer-Verlag, Berlin, New York, 1976.
[15] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1979.
[16] W. I. ZANGWILL AND C. B. GARCIA, *Equilibrium programming: the path-following approach and dynamics*, Math. Programming, 21 (1981), pp. 262–289.
[17] S. NGUYEN AND C. DUPUIS, *An efficient method for computing traffic equilibria in networks with asymmetric transportation costs*, Transportation Sci., 18 (1984), pp. 185–202.
[18] J. M. DANSKIN, *The theory of max-min, with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–664.

[19] J. S. PANG, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Math. Oper. Res., 12 (1987), pp. 474–484.

[20] P. MARCOTTE AND J. GUELAT, *Adaptation of a modified Newton method for solving the asymmetric traffic equilibrium problem*, Transportation Sci., 22 (1988), pp. 112–124.

[21] J.-P. DUSSAULT AND P. MARCOTTE, *Conditions de régularité géométrique pour les inéquations variationnelles*, RAIRO Rech. Opér., 23 (1988), pp. 1–16.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1279-1292, November 1989

003

THE PARTIALLY OBSERVED STOCHASTIC MINIMUM PRINCIPLE*

JOHN S. BARAS†, ROBERT J. ELLIOTT‡, AND MICHAEL KOHLMANN§

Abstract. Using stochastic flows and the generalized differentiation formula of Bismut and Kunita, the change in cost due to a strong variation of an optimal control is explicitly calculated. Differentiating this expression gives a minimum principle in both the partially observed and stochastic open loop situations. In the latter case the equation satisfied by the adjoint process is obtained by applying a martingale representation result.

Key words. stochastic control, minimum principle, adjoint process, stochastic flow

AMS(MOS) subject classification. 93E20

1. Introduction. Various proofs have been given of the minimum principle satisfied by an optimal control in a partially observed stochastic control problem. See, for example, the papers by Bensoussan [1], Elliott [8], Haussmann [11], and the recent paper [14] by Haussmann in which the adjoint process is identified. The simple case of a partially observed Markov chain is discussed in the University of Maryland lecture notes [9] of Elliott.

In this article we show that the minimum principle for a partially observed diffusion can be obtained by differentiating the statement that a control u^* is optimal. The results of Bismut [5], [6] and Kunita [16] on stochastic flows enable us to compute in an easy and explicit way the change in the cost due to a "strong variation" of an optimal control. The only technical difficulty is the justification of the differentiation. As we wished to exhibit the simplification obtained by using the ideas of stochastic flows, the result is not proved under the weakest possible hypotheses. In § 6, stochastic open loop controls are considered and a similar minimum principle with an explicit adjoint process is derived in § 7. If the optimal control is Markov, the equation satisfied by the adjoint process is obtained in § 8 using the martingale representation result of [10]. This simplifies the proof of Haussmann [12]. Finally in § 9 it is pointed out how Bensoussan's minimum principle [2] follows from our result if the drift coefficient is differentiable in the control variable.

2. Dynamics. Suppose the state of the system is described by a stochastic differential equation

$$\begin{aligned} &d\xi_t = f(t, \xi_t, u)\,dt + g(t, \xi_t)\,dw_t, \\ &\xi_t \in R^d, \quad \xi_0 = x_0, \quad 0 \leqq t \leqq T. \end{aligned} \tag{2.1}$$

The control parameter u will take values in a compact subset U of some Euclidean space R^k. We shall make the following assumptions:

* Received by the editors May 25, 1987; accepted for publication (in revised form) January 8, 1989.

† Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742. The work of this author was partially supported by U.S. Army contract DAAL03-86-C-0014 and by National Science Foundation grant CDR-85-00108.

‡ Department of Statistics and Applied Probability, University of Alberta, Edmonton, Alberta, Canada T6G 2G1. The work of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A-7964, and partially supported by U.S. Air Force Office of Scientific Research grant AFOSR-86-0332, the European Office of Aerospace Research and Development, London, United Kingdom, and by the U.S. Office of Naval Research under grant N00014-86-K-0122 through the Systems Research Center.

§ Fakultat fur Wirtschaftswissenschaften und Statistik, Universitat Konstanz, Postfach 5560, D-7750, Federal Republic of Germany. The work of this author was supported by the Natural Sciences and Engineering Research Council of Canada under grant A-7964.

(A_1) x_0 is given; if x_0 is a random variable and P_0 its distribution, the situation when $\int |x|^q P_0(dx) < \infty$ for some $q > n+1$ can be treated, as in [14], by including an extra integration with respect to P_0.

(A_2) $f:[0, T]\times R^d \times U \to R^d$ is Borel measurable, continuous in u for each (t, x), continuously differentiable in x and for some constant K_1, $(1+|x|)^{-1}|f(t, x, u)|+|f_x(t, x, u)| \leqq K_1$.

(A_3) $g:[0, T]\times R^d \to R^d \otimes R^n$ is a matrix-valued function, Borel measurable, continuously differentiable in x, and for some constant K_2, $|g(t, x)|+|g_x(t, x)| \leqq K_2$.

The observation process is given by

$$dy_t = h(\xi_t)\,dt + dv_t, \qquad y_t \in R^m, \quad y_0 = 0, \quad 0 \leqq t \leqq T. \tag{2.2}$$

In the above equations $w=(w^1, \cdots, w^n)$ and $v=(v^1, \cdots, v^d)$ are independent Brownian motions. We also assume the following:

(A_4) $h: R^d \to R^m$ is Borel measurable, continuously differentiable in x, and for some constant K_3, $|h(t, x)|+|h_x(t, x)| \leqq K_3$.

Remark 2.1. These hypotheses can be weakened. For example, in (A_4), h can be allowed linear growth in x. Because g is bounded, a delicate argument then implies the exponential Z of (2.3) is in some L^p space, $1 < p < \infty$. (See, for example, Theorem 2.2 of [13].) However, when h is bounded, Z is in all the L^p spaces (see Lemma 2.3). Also, if we require f to have linear growth in u, then the set of control values U can be unbounded as in [14]. Our objective, however, is not the greatest generality but is to demonstrate the simplicity of the techniques of stochastic flows.

Let $\hat{P}$ denote Wiener measure on $C([0, T], R^n)$ and μ denote Wiener measure on $C([0, T], R^m)$. Consider the space $\Omega = C([0, T], R^n)\times C([0, T], R^m)$ with coordinate functions (w_t, y_t) and define Wiener measure P on Ω by

$$P(dw, dy) = \hat{P}(dw)\mu(dy).$$

DEFINITION 2.2. Write $Y=\{Y_t\}$ for the right continuous complete filtration on $C([0, T], R^m)$ generated by $Y_t^0 = \sigma\{y_s: s \leqq t\}$. The set of admissible control functions $\underline{U}$ will be the Y-predictable functions on $[0, T]\times C([0, T], R^m)$ with values in U.

For $u \in \underline{U}$ and $x \in R^d$ write $\xi^u_{s,t}(x)$ for the strong solution of (2.1) corresponding to control u, and with $\xi^u_{s,s}(x) = x$. Write

$$Z^u_{s,t}(x) = \exp\left(\int_s^t h(\xi^u_{s,r}(x))'\,dy_r - \frac{1}{2}\int_s^t h(\xi^u_{s,r}(x))^2\,dr\right) \tag{2.3}$$

and define a new probability measure P^u on Ω by $dP^u/dP = Z^u_{0,T}(x_0)$. Then under P^u, $(\xi^u_{0,t}(x_0), y_t)$ is a solution of (2.1) and (2.2), that is, $\xi^u_{0,t}(x_0)$ remains a strong solution of (2.1) and there is an independent Brownian motion v such that y_t satisfies (2.2). A version of Z defined for every trajectory y of the observation process is obtained by integrating by parts the stochastic integral in (2.3).

LEMMA 2.3. *Under hypothesis* (A_4) *for* $t \leqq T$,

$$E[(Z^u_{0,t}(x_0))^p] < \infty \quad \textit{for all } u \in \underline{U} \textit{ and all } p, \quad 1 \leqq p < \infty.$$

Proof.

$$Z^u_{0,t}(x_0) = 1 + \int_0^t Z^u_{0,r}(x_0) h(\xi^u_{0,r}(x_0))'\,dy_r.$$

Therefore, for any p there is a constant C_p such that

$$E[(Z^u_{0,t}(x_0))^p] \leqq C_p\left[1+E\left(\int_0^t (Z^u_{0,r}(x_0))^2 h(\xi^u_{0,r}(x_0))^2\,dr\right)^{p/2}\right].$$

The result follows by Gronwall's inequality.

Cost 2.4. We shall suppose the cost is purely terminal and given by some bounded, continuously differentiable function

$$c(\xi^u_{0,T}(x_0)),$$

which has bounded derivatives. Then the expected cost, if control $u \in \underline{U}$ is used, is

$$J(u) = E_u[c(\xi^u_{0,T}(x_0))].$$

In terms of P, under which y_t is always a Brownian motion, this is

$$J(u) = E[Z^u_{0,T}(x_0)c(\xi^u_{0,T}(x_0))]. \tag{2.4}$$

3. Stochastic flows. For $u \in \underline{U}$ write

$$\xi^u_{s,t}(x) = x + \int_s^t f(r, \xi^u_{s,r}(x), u_r)\,dr + \int_s^t g(r, \xi^u_{s,r}(x))\,dw_r \tag{3.1}$$

for the solution of (2.1) over the time interval $[s, t]$ with initial condition $\xi^u_{s,s}(x) = x$. In the sequel we wish to discuss the behavior of (3.1) for each trajectory y of the observation process. We have already noted that there is a version of Z defined for every y. The results of Bismut [5] and Kunita [16] extend easily and show the map

$$\xi^u_{s,t}: R^d \to R^d$$

is, almost surely, for each $y \in C([0, T], R^m)$ a diffeomorphism. Bismut [5] initially gives proofs when the coefficients f and g are bounded, but points out that a stopping time argument extends the results to when, for example, the coefficients have linear growth.

Write $\|\xi^u(x_0)\|_t = \sup_{0\leqq x\leqq t} |\xi^u_{0,s}(x_0)|$. Then, as in Lemma 2.1 of [13], for any p, $1 \leqq p < \infty$, using Gronwall's and Jensen's inequalities,

$$\|\xi^u(x_0)\|^p_T \leqq C\left(1+|x_0|^p + \left|\int_0^t g(r, \xi^u_{0,r}(x_0))\,dw_r\right|^p\right)$$

almost surely, for some constant C.

Therefore, using Burkholder's inequality and hypothesis (A_3), $\|\xi^u(x_0)\|_T$ is in L^p for all p, $1 \leqq p < \infty$.

Suppose $u^* \in \underline{U}$ is an optimal control; then $J(u^*) \leqq J(u)$ for any other $u \in \underline{U}$. Write $\xi^*_{s,t}(\cdot)$ for $\xi^{u^*}_{s,t}(\cdot)$. The derivative $\partial\xi^*_{s,t}(x)/\partial x$ is the matrix solution C_t of the equation for $s \leqq t$,

$$dC_t = f_x(t, \xi^*_{s,t}(x), u^*)C_t\,dt + \sum_{i=1}^n g^{(i)}_x(t, \xi^*_{s,t}(x))C_t\,dw^i_t \quad \text{with } C_s = I. \tag{3.2}$$

Here I is the $n \times n$ identity matrix and $g^{(i)}$ is the ith column of g. From hypotheses (A_2) and (A_3), f_x and g_x are bounded. When we write $\|C\|_t = \sup_{0\leqq s\leqq t} |C_s|$, an application of Gronwall's, Jensen's, and Burkholder's inequalities again implies $\|C\|_T$ is in

L^p for all p, $1 \leqq p < \infty$. Consider the related matrix-valued stochastic differential equation

$$
\begin{aligned}
D_t = I - \int_s^t D_r f_x(r, \xi^*_{s,r}(x), u^*_r)' \, dr - \sum_{i=1}^n \int_s^t D_r g_x^{(i)}(r, \xi^*_{s,r}(x))' \, dw_r^i \\
+ \sum_{i=1}^n \int_s^t D_r (g_x^{(i)}(r, \xi^*_{s,r}(x))')^2 \, dr.
\end{aligned} \tag{3.3}
$$

Then it can be checked that $D_t C_t = I$ for $t \geqq s$, so that D_t is the inverse of the Jacobian, that is, $D_t = (\partial \xi^*_{s,t}(x)/\partial x)^{-1}$. Again, because f_x and g_x are bounded we have that $\|D\|_t$ is in every L^p, $1 \leqq p < \infty$.

For a d-dimensional semimartingale z_t Bismut [5] shows that $\xi^*_{s,t}(z_t)$ is well-defined and gives the semimartingale representation of this process. In fact if $z_t = z_s + A_t + \sum_{i=1}^n \int_s^t H_i \, dw_r^i$ is a d-dimensional semimartingale, Bismut's formula states that

$$
\begin{aligned}
\xi^*_{s,t}(z_t) = z_s + \int_s^t \left(f(r, \xi^*_{s,r}(z_r), u^*_r) + \sum_{i=1}^n g_x^{(i)}(r, \xi^*_{s,r}(z_r), u^*_r) \frac{\partial \xi^*_{s,r}}{\partial x}(z_r) H_i \right. \\
\left. + \frac{1}{2} \sum_{i=1}^n \frac{\partial^2 \xi^*_{s,r}(z_r)}{\partial x^2}(H_i, H_i) \right) dr \\
+ \int_s^t \frac{\partial \xi^*_{s,r}(z_r)}{\partial x} \, dA_r + \sum_{i=1}^n \int_s^t \left(g^{(i)}(r, \xi^*_{s,r}(z_r)) + \frac{\partial \xi^*_{s,r}}{\partial x}(z_r) H_i \right) dw_r^i.
\end{aligned} \tag{3.4}
$$

DEFINITION 3.1. We shall consider perturbations of the optimal control u^* of the following kind. For $s \in [0, T)$, $h > 0$ such that $0 \leqq s < s + h \leqq T$, for any other admissible control $\tilde{u} \in \mathcal{U}$ and $A \in Y_s$ define a strong variation of u^* by

$$
u(t, w) = \begin{cases} u^*(t, w) & \text{if } (t, w) \notin [s, s+h] \times A, \\ \tilde{u}(t, w) & \text{if } (t, w) \in [s, s+h] \times A. \end{cases}
$$

Applying (3.4) as in Theorem 5.1 of [7], we have the following result.

THEOREM 3.2. *For the perturbation u of the optimal control u^* consider the process*

$$
z_t = x + \int_s^t \left(\frac{\partial \xi^*_{s,r}(z_r)}{\partial x} \right)^{-1} (f(r, \xi^*_{s,r}(z_r), u_r) - f(r, \xi^*_{s,r}(z_r), u^*_r)) \, dr. \tag{3.5}
$$

Then the process $\xi^*_{s,t}(z_t)$ is indistinguishable from $\xi^u_{s,t}(x)$.

Proof. Note that the equation defining z_t involves only an integral in time; there is no martingale term, so to apply (3.4) we have $H_i = 0$ for all i. Therefore, from (3.4)

$$
\begin{aligned}
\xi^*_{s,t}(z_t) = x + \int_s^t f(r, \xi^*_{s,r}(z_r), u^*_r) \, dr \\
+ \int_s^t \left(\frac{\partial \xi^*_{s,r}(z_r)}{\partial x} \right) \left(\frac{\partial \xi^*_{s,r}(z_r)}{\partial x} \right)^{-1} (f(r, \xi^*_{s,r}(z_r), u_r) - f(r, \xi^*_{s,r}(z_r), u^*_r)) \, dr \\
+ \int_s^t g(r, \xi^*_{s,r}(z_r)) \, dw_r.
\end{aligned}
$$

However, the solution of (3.1) is unique so

$$
\xi^*_{s,t}(z_t) = \xi^u_{s,t}(x).
$$

Remark 3.3. Note that the perturbation $u(t)$ equals $u^*(t)$ if $t > s + h$ so $z_t = z_{s+h}$ if $t > s + h$ and

$$
\xi^*_{s,t}(z_t) = \xi^*_{s,t}(z_{s+h}) = \xi^*_{s+h,t}(\xi^u_{s,s+h}(x)).
$$

4. Augmented flows. Consider the augmented flow that includes as an extra coordinate the stochastic exponential $Z^*_{s,t}$ with a "variable" initial condition $z \in R$ for $Z^*_{s,s}(\cdot)$. That is, consider the $(d+1)$-dimensional system given by

$$\xi^*_{s,t}(x) = x + \int_s^t f(r, \xi^*_{s,r}(x), u^*_r)\, dr + \int_s^t g(r, \xi^*_{s,r}(x))\, dw_r,$$

$$Z^*_{s,t}(x, z) = z + \int_s^t Z^*_{s,r}(x, z) h(\xi^*_{s,r}(x))'\, dy_r.$$

Therefore, from the first equation in the proof of Lemma 2.3 we have

$$\begin{aligned} Z^*_{s,t}(x, z) &= z Z^*_{s,t}(x) \\ &= z \exp\left(\int_s^t h(\xi^*_{s,r}(x))'\, dy_r - \frac{1}{2} \int_s^t h(\xi^*_{s,r}(x))^2\, dr \right) \end{aligned}$$

and we see there is a version of the enlarged system defined for each trajectory y by integrating by parts the stochastic integral. The augmented map $(x, z) \to (\xi^*_{s,t}(x), Z^*_{s,t}(x, z))$ is then almost surely a diffeomorphism of R^{d+1}. Note that $\partial \xi^*_{s,t}(x)/\partial z = 0$, $\partial f/\partial z = 0$ and $\partial g/\partial z = 0$. The Jacobian of this augmented map is, therefore, represented by the matrix

$$\tilde{C}_t = \begin{pmatrix} \partial \xi^*_{s,t}(x)/\partial x & 0 \\ \partial Z^*_{s,t}(x, z)/\partial x & \partial Z^*_{s,t}(x, z)/\partial z \end{pmatrix},$$

and for $1 \leqq i \leqq d$ as in (3.2)

$$\begin{aligned} \frac{\partial Z^*_{s,t}(x, z)}{\partial x_i} = \sum_{j=1}^m \int_s^t \Bigg(& Z^*_{s,r}(x, z) \frac{\partial h^j(\xi^*_{s,r}(x))}{\partial \xi_k} \cdot \frac{\partial \xi^*_{k,s,r}(x)}{\partial x_i} \\ & + h^j(\xi^*_{s,r}(x)) \frac{dZ^*_{s,r}(x, z)}{\partial x_i} \Bigg)\, dy^j_r. \end{aligned} \tag{4.1}$$

(Here the double index k is summed from 1 to n.)

We shall be interested in the solution of this differential system (4.1) only in the situation when $z = 1$, so we shall write $Z^*_{s,t}(x)$ for $Z^*_{s,t}(x, 1)$. The following result is motivated by formally differentiating the exponential formula for $Z^*_{s,t}(x)$.

LEMMA 4.1.

$$\frac{\partial Z^*_{s,t}(x)}{\partial x} = Z^*_{s,t}(x) \left(\int_s^t h_x(\xi^*_{s,r}(x)) \cdot \frac{\partial \xi^*_{s,r}(x)}{\partial x} \cdot dv_r \right)$$

where $v = (v^1, \cdots, v^n)$ is the Brownian motion in the observation process.

Proof. From (4.1) we see $\partial Z^*_{s,t}(x)/\partial x$ is the solution of the stochastic differential equation

$$\frac{\partial Z^*_{s,t}(x)}{\partial x} = \int_s^t \left(\frac{\partial Z^*_{s,r}(x)}{\partial x} h'(\xi^*_{s,r}(x)) + Z^*_{s,r}(x) h_x(\xi^*_{s,r}(x)) \frac{\partial \xi^*_{s,r}(x)}{\partial x} \right) dy_r. \tag{4.2}$$

Write

$$L_{s,t}(x) = Z^*_{s,t}(x) \left(\int_s^t h_x \cdot \frac{\partial \xi^*_{s,r}}{\partial x} \cdot dv_r \right)$$

where

$$dy_r = h(\xi^*_{s,t}(x))\, dt + dv_t.$$

Because

$$Z^*_{s,t}(x) = 1 + \int_s^t Z^*_{s,r}(x) h'(\xi^*_{s,r}(x))\, dy_r$$

the product rule gives

$$\begin{aligned} L_{s,t}(x) &= \int_s^t Z^*_{s,r}(x) h_x \cdot \frac{\partial \xi^*_{s,r}}{\partial x}\, dv_r + \int_s^t \left(\int_s^r h_x \cdot \frac{\partial \xi^*_{s,\sigma}}{\partial x} \cdot dv_\sigma \right) Z^*_{s,r}(x) h'(\xi^*_{s,r}(x))\, dy_r \\ &\quad + \int_s^t Z^*_{s,r}(x) h'(\xi^*_{s,r}(x)) \cdot h_x \cdot \frac{\partial \xi^*_{s,r}}{\partial x}\, dr \\ &= \int_s^t L_{s,r}(x) h'(\xi^*_{s,r}(x))\, dy_r + \int_s^t Z^*_{s,r}(x) h_x \cdot \frac{\partial \xi^*_{s,r}}{\partial x} \cdot dy_r. \end{aligned}$$

Therefore, $L_{s,t}(x)$ is also a solution of (4.2), so by uniqueness

$$L_{s,t}(x) = \frac{\partial Z^*_{s,t}(x)}{\partial x}.$$

Remark 4.2. As noted at the beginning of this section we can consider the augmented flow

$$(x, z) \to (\xi^*_{s,t}(x), Z^*_{s,t}(x, z)) \quad \text{for } x \in R^d,\ z \in R,$$

and we are only interested in the situation when $z = 1$, so we write $Z^*_{s,t}(x)$.

LEMMA 4.3. *$Z^*_{s,t}(z_t) = Z^u_{s,t}(x)$ where z_t is the semimartingale defined in* (3.6).

Proof. $Z^u_{s,t}(x)$ is the process uniquely defined by

$$Z^u_{s,t}(x) = 1 + \int_s^t Z^u_{s,r}(x) h'(\xi^u_{s,r}(x))\, dy_r. \tag{4.3}$$

Consider an augmented $(d+1)$-dimensional version of (3.5) defining a semimartingale $\bar{z}_t = (z_t, 1)$, so the additional component is always identically one. Then applying (3.4) to the new component of the augmented process, we have

$$\begin{aligned} Z^*_{s,r}(z_r) &= 1 + \int_s^t Z^*_{s,r}(z_r) h'(\xi^*_{s,r}(z_r))\, dy_r \\ &= 1 + \int_s^t Z^*_{s,r}(z_r) h'(\xi^u_{s,r}(x))\, dy_r \end{aligned}$$

by Theorem 3.2. However, (4.3) has a unique solution so $Z^*_{s,t}(z_t) = Z^u_{s,t}(x)$.

Remark 4.4. Note that for $t > s + h$

$$Z^*_{s,t}(z_t) = Z^*_{s,t}(z_{s+h}).$$

5. The minimum principle. Control u will be the perturbation of the optimal control u^* as in Definition 3.1. We shall write $x = \xi^*_{0,s}(x_0)$. Then the minimum cost is

$$\begin{aligned} J(u^*) &= E[Z^*_{0,T}(x_0) c(\xi^*_{0,T}(x_0))] \\ &= E[Z^*_{0,s}(x_0) Z^*_{s,T}(x) c(\xi^*_{s,T}(x))]. \end{aligned}$$

The cost corresponding to the perturbed control u is

$$\begin{aligned} J(u) &= E[Z^*_{0,s}(x_0) Z^u_{s,T}(x) c(\xi^u_{s,T}(x))] \\ &= E[Z^*_{0,s}(x_0) Z^*_{s,T}(z_{s+h}) c(\xi^*_{s,T}(z_{s+h}))] \end{aligned}$$

by Theorem 3.2 and Lemma 4.3. Now $Z^*_{s,T}(\cdot)$ and $c(\xi^*_{s,T}(\cdot))$ are almost surely differentiable with continuous derivatives and z_t, given by (3.5), is absolutely continuous. Therefore,

$$J(u)-J(u^*)=E[Z^*_{0,s}(x_0)(Z^*_{s,T}(z_{s+h})c(\xi^*_{s,T}(z_{s+h}))-Z^*_{s,T}(x)c(\xi^*_{s,T}(x)))]$$
$$=E\left[\int_s^{s+h}\Gamma(s,z_r)(f(r,\xi^*_{s,r}(z_r),u^*_r)-f(r,\xi^*_{s,r}(x),u^*_r))\,dr\right]$$

where by Lemma 4.1

$$\Gamma(s,z_r)=Z^*_{0,s}(x_0)Z^*_{s,T}(z_r)\left\{c_\xi(\xi^*_{s,T}(z_r))\frac{\partial\xi^*_{s,T}(z_r)}{\partial x}\right.$$
$$\left.+c(\xi^*_{s,T}(z_r))\left(\int_s^T h_\xi(\xi^*_{s,\sigma}(z_r))\frac{\partial\xi^*_{s,\sigma}}{\partial x}(z_r)\,dv_\sigma\right)\right\}\left(\frac{\partial\xi^*_{s,r}}{\partial x}(z_r)\right)^{-1}.$$

Note that this expression gives an explicit formula for the change in the cost resulting from a variation in the optimal control. The only remaining problem is to justify differentiating the right-hand side.

From Lemma 2.3, Z is in every L^p space, $1\leqq p<\infty$, and from the remarks at the beginning of § 3, $C_T=\partial\xi^*_{s,T}/\partial x$ and $D_T=(\partial\xi^*_{s,T}/\partial x)^{-1}$ are in every L^p space, $1\leqq p<\infty$. Consequently, Γ is in every L^p space, $1\leqq p<\infty$.

Therefore,

$$J(u)-J(u^*)=\int_s^{s+h}E[(\Gamma(s,z_r)-\Gamma(s,x))(f(r,\xi^*_{s,r}(z_r),u_r)-f(r,\xi^*_{s,r}(z_r),u^*_r))]\,dr$$
$$+\int_s^{s+h}E[(\Gamma(s,x)-\Gamma(r,x))(f(r,\xi^*_{s,r}(z_r),u_r)-f(r,\xi^*_{s,r}(z_r),u^*_r))]\,dr$$
$$+\int_s^{s+h}E[\Gamma(r,x)(f(r,\xi^*_{s,r}(z_r),u_r)-f(r,\xi^*_{s,r}(z_r),u^*_r)$$
$$-f(r,\xi^*_{s,r}(x),u_r)+f(r,\xi^*_{s,r}(x),u^*_r))]\,dr$$
$$+\int_s^{s+h}E[\Gamma(r,x)(f(r,\xi^*_{0,r}(x_0),u_r)-f(r,\xi^*_{0,r}(x_0),u^*_r))]\,dr$$
$$=I_1(h)+I_2(h)+I_3(h)+I_4(h),\quad\text{say.}$$

Now,

$$|I_1(h)|\leqq K_1\int_s^{s+h}E[|\Gamma(s,z_r)-\Gamma(s,x)|(1+\|\xi^u(x_0)\|_{s+h})]\,dr$$
$$\leqq K_1h\sup_{s\leqq r\leqq s+h}E[|\Gamma(s,z_r)-\Gamma(s,x)|(1+\|\xi^u(x_0)\|_{s+h})],$$
$$|I_2(h)|\leqq K_2\int_s^{s+h}E[|\Gamma(s,x)-\Gamma(r,s)|(1+\|\xi^u(x_0)\|_{s+h})]\,dr$$
$$\leqq K_2h\sup_{s\leqq r\leqq s+h}E[|\Gamma(s,x)-\Gamma(r,x)|(1+\|\xi^u(x_0)\|_{s+h})],$$
$$|I_3(h)|\leqq K_3\int_s^{s+h}E[|\Gamma(r,x)|\,\|\xi^*_{s,r}(z_r)-\xi^*_{s,r}(x)\|]\,dr$$
$$\leqq K_3h\sup_{s\leqq r\leqq s+h}E[|\Gamma(r,x)|\,\|\xi^u_{s,\cdot}(x)-\xi^*_{s,\cdot}(x)\|_{s+h}].$$

The differences $|\Gamma(s, z_r)-\Gamma(s, x)|$, $|\Gamma(s, x)-\Gamma(r, x)|$ and $\|\xi^u_{s,\cdot}(x)-\xi^*_{s,\cdot}(x)\|_{s+h}$ are all uniformly bounded in some L^p, $p \geqq 1$, and

$$\lim_{r\to s} |\Gamma(s, z_r)-\Gamma(s, x)| = 0 \quad \text{a.s.},$$

$$\lim_{r\to s} |\Gamma(s, x)-\Gamma(r, x)| = 0 \quad \text{a.s.},$$

$$\lim_{h\to 0} \|\xi^u_{s,\cdot}(x)-\xi^*_{s,\cdot}(x)\|_{s+h} = 0.$$

Therefore,

$$\lim_{r\to s} \|\Gamma(s, z_r)-\Gamma(s, x)\|_p = 0,$$

$$\lim_{r\to s} \|\Gamma(s, x)-\Gamma(r, x)\|_p = 0, \quad \text{and}$$

$$\lim_{h\to 0} \|(\|\xi^u_{s,\cdot}(x)-\xi^*_{s,\cdot}(x)\|_{s+h})\|_p = 0 \quad \text{for some } p.$$

Consequently, $\lim_{h\to 0} h^{-1} I_k(h) = 0$, for $k = 1, 2, 3$.

The only remaining problem concerns the differentiability of

$$I_4(h) = \int_s^{s+h} E[\Gamma(r, x)(f(r, \xi^*_{0,r}(x_0), u_r)-f(r, \xi^*_{0,r}(x_0), u^*_r))]\, dr.$$

The integrand is almost surely in $L^1([0, T])$ so $\lim_{h\to 0} h^{-1} I_4(h)$ exists for almost every $s \in [0, T]$. However, the set of times $\{s\}$ where the limit may not exist might depend on the control u. Consequently we must restrict the perturbations u of the optimal control u^* to perturbations from a countable dense set of controls. In fact:

(1) Because the trajectories are, almost surely, continuous, Y_ρ is countably generated by sets $\{A_{i\rho}\}$, $i = 1, 2, \cdots$ for any rational number $\rho \in [0, T]$. Consequently, Y_t is countably generated by the sets $\{A_{i\rho}\}$, $\rho \leqq t$.

(2) Let G_t denote the set of measurable functions from (Ω, Y_t) to $U \subset R^k$. (If $u \in \underline{U}$ then $u(t, w) \in G_t$.) Using the L^1-norm, as in [8], there is a countable dense subset $H_\rho = \{u_{j\rho}\}$ of G_ρ, for rational $\rho \in [0, T]$. If $H_t = \bigcup_{\rho \leqq t} H_\rho$ then H_t is a countable dense subset of G_t. If $u_{j\rho} \in H_\rho$ then, as a function constant in time, $u_{j\rho}$ can be considered as an admissible control over the time interval $[t, T]$ for $t \geqq \rho$.

(3) The countable family of perturbations is obtained by considering sets $A_{i\rho} \in Y_t$, functions $u_{j\rho} \in H_t$, where $\rho \leqq t$, and defining as in (3.1) the following:

$$u^*_{j\rho}(s, w) = \begin{cases} u^*(s, w) & \text{if } (s, w) \notin [t, T] \times A_{i\rho}, \\ u_{j\rho}(s, w) & \text{if } (s, w) \in [t, T] \times A_{i\rho}. \end{cases}$$

Then for each i, j, ρ

$$\lim_{h\to 0} h^{-1} \int_s^{s+h} E[\Gamma(r, x)(f(r, \xi^*_{0,r}(x_0), u^*_{j\rho})-f(r, \xi^*_{0,r}(x_0), u^*))]\, dr \tag{5.1}$$

exists and equals

$$E[\Gamma(s, x)(f(s, \xi^*_{0,s}(x_0), u_{j\rho})-f(s, \xi^*_{0,s}(x_0), u^*)) I_{A_{i\rho}}]$$

for almost all $s \in [0, T]$. Therefore, considering this perturbation we have

$$\lim_{h\to 0} h^{-1}(J(u^*_{j\rho})-J(u^*)) = E[\Gamma(s, x)(f(s, \xi^*_{0,s}(x_0), u_{j\rho})-f(s, \xi^*_{0,s}(x_0), u^*)) I_{A_{i\rho}}]$$

$$\geqq 0 \quad \text{for almost all } s \in [0, T].$$

Consequently there is a set $S \subset [0, T]$ of zero Lebesgue measure such that, if $s \notin S$, the limit in (5.1) exists for all i, j, ρ, and gives

$$E[\Gamma(s, x)(f(s, \xi^*_{0,s}(x_0), u_{j\rho}) - f(s, \xi^*_{0,s}(x_0), u^*))I_{A_{i\rho}}] \geqq 0.$$

Using the monotone class theorem, and approximating an arbitrary admissible control $u \in \underline{U}$, we can deduce that if $s \notin S$, then

$$E[\Gamma(s, x)(f(s, \xi^*_{0,s}(x_0), u) - f(s, \xi^*_{0,s}(x_0), u^*))I_A] \geqq 0 \quad \text{for any } A \in Y_s. \tag{5.2}$$

Write

$$p_s(x) = E^*\left[c_\xi(\xi^*_{0,T}(x_0)) \frac{\partial \xi^*_{s,T}(x)}{\partial x} + c(\xi^*_{0,T}(x_0))\left(\int_s^T h_\xi(\xi^*_{0,\sigma}(x_0)) \frac{\partial \xi^*_{s,\sigma}(x)}{\partial x}\, dv_\sigma \right) \middle| \, Y_s \vee \{x\} \right]$$

where, as before, $x = \xi^*_{0,s}(x_0)$ and E^* denotes expectation under $P^* = P^{u^*}$. Then $p_s(x)$ is the co-state variable and we have in (5.2) proved the following "conditional" minimum principle.

THEOREM 5.1. *If $u^* \in \underline{U}$ is an optimal control there is a set $S \subset [0, T]$ of zero Lebesgue measure such that if $s \notin S$*

$$E^*[p_s(x)f(s, x, u^*) \mid Y_s] \geqq E^*[p_s(x)f(s, x, u) \mid Y_s] \quad a.s.$$

That is, the optimal control u^* almost surely minimizes the conditional Hamiltonian and the adjoint variable is $p_s(x)$.

6. Stochastic open loop controls. We shall again suppose the state of the system is described by a stochastic differential equation

$$d\xi_t = f(t, \xi_t, u)\, dt + g(t, \xi_t)\, dw_t, \qquad \xi_t \in R^d, \quad \xi_0 = x_0, \quad 0 \leqq t \leqq T \tag{6.1}$$

where x_0, f, and g satisfy the same assumptions A_1, A_2, and A_3 as in § 2.

Suppose $w = (w^1, \cdots, w^n)$ is an n-dimensional Brownian motion on a probability space (Ω, F, P), with a right continuous complete filtration $\{F_t\}$, $0 \leqq t \leqq T$. Rather than controls depending on some observation process y we now consider controls that depend on the "noise process" w. These are sometimes called "stochastic open loop" controls [4].

DEFINITION 6.1. The set of admissible controls $\underline{V}$ will be the F_t-predictable functions on $[0, T] \times \Omega$ with values in a compact subset V of some Euclidean space R^k.

Remark 6.2. For each $u \in \underline{V}$ there is, therefore, a strong solution of (6.1) and we shall write $\xi^u_{s,t}(x)$ for the solution trajectory given by

$$\xi^u_{s,t}(x) = x + \int_s^t f(r, \xi^u_{s,r}(x), u_r)\, dr + \int_s^t g(r, \xi^u_{s,r}(x))\, dw_r. \tag{6.2}$$

Again, because u is a (predictable) parameter the results of [2], [5], or [16] extend to this situation, so the derivative $\partial \xi^u_{s,t}/\partial x(x) = C^u_{s,t}$ exists and is the solution of

$$C^u_{s,t} = I + \int_s^t f_\xi(r, \xi^u_{s,r}(x), u_r) C^u_{s,r}\, dr + \sum_{k=1}^n \int_s^t g^{(k)}_\xi(r, \xi^u_{s,r}(x)) C^u_{s,r}\, dw^k_r. \tag{6.3}$$

Suppose $D^u_{s,t}$ is the matrix-valued process defined by

$$\begin{aligned} D^u_{s,r} = I &- \int_s^t D^u_{s,t}\left(f_\xi(r, \xi^u_{s,r}(x), u_r) - \sum_{k=1}^n g^{(k)}_\xi(r, \xi^u_{s,r}(x))^2 \right) dr \\ &- \sum_{k=1}^n \int_s^t D^u_{s,r} g^{(k)}_\xi(r, \xi^u_{s,r}(x))\, dw^k_r. \end{aligned} \tag{6.4}$$

Using the Itô rule as in § 3 we see that $d(D^u_{s,t}C^u_{s,t})=0$ and $D^u_{s,s}C^u_{s,s}=I$, so

$$D^u_{s,t}=(C^u_{s,t})^{-1}.$$

As before, if

$$\|\xi^u(x_0)\|_t=\sup_{0\leqq s\leqq t}|\xi^u_{0,s}(x_0)|,$$

$$\|C^u\|_T=\sup_{0\leqq s\leqq T}|C^u_{0,s}|,\qquad \|D^u\|_T=\sup_{0\leqq s\leqq T}|D^u_{0,s}|,$$

then applications of Gronwall's, Jensen's, and Burkholder's inequalities imply that

$$\|\xi^u(x_0)\|_k,\quad \|C^u\|_T,\quad \text{and}\quad \|D^u\|_T$$

are in L^p for all p, $1\leqq p<\infty$.

Cost 6.3. As in § 2, we shall suppose the cost is purely terminal and given by a bounded C^2 function

$$c(\xi^u_{0,T}(x_0)).$$

Furthermore, we shall assume

$$|c(x)|+|c_x(x)|+|c_{xx}(x)|\leqq K_3(1+|x|^q)$$

for some $q<\infty$.

The expected cost if a control $u\in\underline{V}$ is used, therefore, is

$$J(u)=E[c(\xi^u_{0,T}(x_0))].$$

Suppose there is an optimal control $u^*\in\underline{V}$ so that

$$J(u^*)\leqq J(u)\quad \text{for all } u\in\underline{V}.$$

Notation 6.4. If u^* is an optimal control, write ξ^* for ξ^{u^*}, C^* for C^{u^*}, etc.

DEFINITION 6.5. Consider perturbations of u^* of the following kind. For $s\in[0,T]$, $h>0$ such that $0\leqq s<s+h\leqq T$ and $A\in F_s$ define, for any other $\tilde{u}\in\underline{V}$, a strong variation of u^* by

$$u(t,w)=\begin{cases}u^*(t,w) & \text{if } (t,w)\notin[s,s+h]\times A,\\ \tilde{u}(t,w) & \text{if } (t,w)\in[s,s+h]\times A.\end{cases}$$

The following result is established exactly as Theorem 3.2.

THEOREM 6.6. *For any perturbation u of u^* consider the process*

$$z_r=x+\int_s^t\left(\frac{\partial\xi^*_{s,r}}{\partial x}(z_r)\right)^{-1}(f(r,\xi^*_{s,r}(z_r),u_r)-f(r,\xi^*_{s,r}(z_r),u^*_r))\,dr. \tag{6.5}$$

Then the process $\xi^*_{s,t}(z_t)$ is indistinguishable from $\xi^u_{s,t}(x)$.

Note if $t>s+h$, $\xi^*_{s,t}(z_t)=\xi^*_{s,t}(z_{s+h})=\xi^*_{s+h,t}(\xi^u_{s,s+h}(x))$.

7. An open loop minimum principle. Now

$$\begin{aligned}J(u^*)&=E[c(\xi^*_{0,T}(x_0))]\\&=E[c(\xi^*_{s,T}(x))]\end{aligned}$$

where $x=\xi^*_{0,s}(x_0)$.

Similarly,

$$\begin{aligned}J(u)&=E[c(\xi^u_{0,T}(x_0))]\\&=E[c(\xi^u_{s,T}(x))]\\&=E[c(\xi^*_{s,T}(z_{s+h}))].\end{aligned}$$

Therefore,

$$J(u)-J(u^*)=E[c(\xi^*_{s,T}(z_{s+h}))-c(\xi^*_{s,T}(x))].$$

Because $\xi^*_{s,T}(\cdot)$ is differentiable this is

$$= E\left[\int_s^{s+h} c_\xi(\xi^*_{s,T}(z_r))\frac{\partial \xi^*_{s,T}}{\partial x}(z_r)\cdot\left(\frac{\partial \xi^*_{s,r}}{\partial x}(z_r)\right)^{-1}(f(r,\xi^*_{s,r}(z_r),u_r) - f(r,\xi^*_{s,r}(z_r),u^*_r))\,dr\right]. \tag{7.1}$$

As in § 5, this gives an explicit formula for the change in the cost resulting from a "strong variation" in the optimal stochastic open loop control. It involves a time integration over $[s, s+h]$ and, again, the only remaining problem is to justify the differentiation of the right-hand side of (7.1).

Write

$$\Gamma(s,r,z_r)=c_\xi(\xi^*_{s,T}(z_r))\frac{\partial \xi^*_{s,T}}{\partial x}(z_r)\left(\frac{\partial \xi^*_{s,r}}{\partial x}(z_r)\right)^{-1}$$

and

$$\begin{aligned} p_s(x) &= E\left[c_\xi(\xi^*_{0,T}(x_0))\frac{\partial \xi^*_{s,T}}{\partial x}(x)\,|\,F_s\right] \\ &= E[\Gamma(s,s,x)\,|\,F_s], \end{aligned} \tag{7.2}$$

where, as above, $x=\xi^*_{0,s}(x_0)$.

Then arguments similar to those of § 5—but in fact simpler because Z is not involved—enable us to show that there is a set $S\subset[0,T]$ of zero Lebesque measure such that if $s\notin S$,

$$E[\Gamma(s,s,x)(f(s,\xi^*_{0,s}(x_0),u)-f(s,\xi^*_{0,s}(x_0),u^*))I_A]\geqq 0$$

for any $u\in V$ and $A\in F_s$.

That is, in terms of the adjoint variable $p_s(x)$ we have the following minimum principle for stochastic open loop controls.

THEOREM 7.1. *If $u^*\in V$ is an optimal stochastic open loop control there is a set $S\subset[0,T]$ of zero Lebesgue measure such that if $s\notin S$*

$$p_s(x)f(s,x,u^*)\leqq p_s(x)f(s,x,u)\quad a.s.$$

for all $u\in V$. That is, the optimal control u^* almost surely minimizes the Hamiltonian with adjoint variable $p_s(x)$.

Remark 7.2. Under certain conditions the minimum cost attainable under the stochastic open loop controls is equal to the minimum cost attainable under the Markov feedback controls of the form $u(s,\xi^u_{0,s}(x_0))$. See for example [3], [12]. If u_M is a Markov control, with a corresponding, possibly weak, solution trajectory ξ^{u_M}, then u_M can be considered as a stochastic open loop control $u_M(w)$ by putting

$$u_M(w)=u_M(s,\xi^{u_M}_{0,s}(x_0,w)).$$

This means the control in effect "follows" its original trajectory ξ^{u_M} rather than any new trajectory. That is, the control is similar to the adjoint strategies considered by Krylov [15]. The significance of this is that when we consider variations in the state trajectory ξ, and derivatives of the map $x\to\xi_{s,t}(x)$, the control does not react, and so we do not introduce derivatives in the u variable.

If the optimal control u^* is the Markov, then the process ξ^* is Markov and

$$\begin{aligned} p_s(x) &= E[\Gamma(s, s, x) \mid F_s] \\ &= E[\Gamma(s, s, x) \mid x]. \end{aligned} \tag{7.3}$$

8. The adjoint process. Suppose the optimal stochastic open loop control u^* is Markov. The Jacobian $\partial \xi^*_{s,T}/\partial x$ exists, as does $(\partial \xi^*_{s,T}/\partial x)^{-1}$ and higher derivatives.

THEOREM 8.1. *Suppose the optimal control u^* is Markov. Then*

$$\begin{aligned} p_s(x) = E[c_\xi(\xi^*_{0,T}(x_0))C_{0,t}] &- \int_0^s p_r(\xi^*_{0,r}(x_0)) f_\xi(r, \xi^*_{0,r}(x_0), u^*_r)\, dr \\ &+ \int_0^s p_x(r, \xi^*_{0,r}(x_0)) g(r, \xi^*_{0,r}(x_0))\, dw_r \\ &- \int_0^s p_x(r, \xi^*_{0,r}(x_0)) g(r, \xi^*_{0,r}(x_0)) g_\xi(r, \xi^*_{0,r}(x_0))\, dr. \end{aligned}$$

Proof. Write $f_\xi(r)$ for $f_\xi(r, \xi^*_{0,r}(x_0), u^*_r)$ and $g(r)$ for $g(r, \xi^*_{0,r}(x_0))$, etc. By uniqueness of the solutions to (6.1)

$$\xi^*_{0,T}(x_0) = \xi^*_{s,T}(\xi^*_{0,s}(x_0)) \tag{8.1}$$

so, differentiating,

$$C_{0,T} = C_{s,T} C_{0,s} \tag{8.2}$$

where $C_{0,T} = C^*_{0,T}$, etc. (without the *).

From (7.2) and (7.3)

$$p_s(x) = E[c_\xi(\xi^*_{0,T}(x_0))C_{s,T} \mid F_s],$$

so from (8.2)

$$p_s(x)C_{0,s} = E[c_\xi(\xi^*_{0,T}(x_0))C_{0,T} \mid F_s], \tag{8.3}$$

and this is a $(P, \{F_t\})$ martingale. Write $x = \xi^*_{0,s}(x_0)$, $C = C_{0,s}$. From the martingale representation result [10], the integrand in the representation of $p_s(x)C$ as a stochastic integral is obtained by the Itô rule, noting that only the stochastic integral terms will appear. These involve the derivatives in x and C. In fact, by considering the system $\bar{\xi}_{0,t}$ with components $\xi^*_{0,t}$ and $C_{0,t}$ and any real C^2 function Φ, the martingale

$$\begin{aligned} M_s &= E[\Phi(\bar{\xi}_{0,T} \mid F_s] = E[\Phi(\bar{\xi}_{0,T}) \mid x, C] = V(s, x, C) \\ &= V(0, x_0, I) + \int_0^s V_x(r, \xi^*_{0,r}(x_0), C_{0,r}) g(r)\, dw_r \\ &\quad + \sum_{k=1}^n \int_0^s V_C(r, \xi^*_{0,r}(x_0), C_{0,r}) g^{(k)}_\xi(r) C_{0,r}\, dw^k_r. \end{aligned}$$

Therefore, for the vector martingale (8.3)

$$\begin{aligned} p_s(x)C = E[c_\xi(\xi_{0,T}(x_0))C_{0,T}] &+ \int_0^s p_x(r, \xi^*_{0,r}(x_0)) g(r)\, dw_r C_{0,r} \\ &+ \sum_{k=1}^n \int_0^s p_r(\xi^*_{0,r}(x_0)) g^{(k)}_\xi(r) C_{0,r}\, dw^k_r. \end{aligned} \tag{8.4}$$

Recall that $D_{0,s} = C^{-1}$, so forming the product of (6.4) and (8.4) by using the Itô rule, we have

$$
\begin{aligned}
p_s(x) &= (p_x(x)C)D_{0,s} \\
&= E[c_\xi(\xi^*_{0,T}(x_0))C_{0,T}] - \int_0^s p_r(\xi^*_{0,r}(x_0))f_\xi(r)\,dr \\
&\quad - \sum_{k=1}^n \int_0^s p_r(\xi^*_{0,r}(x_0))g^{(k)}_\xi(r)\,dw^k_r + \sum_{k=1}^n \int_0^s p_r(\xi^*_{0,r}(x_0))(g^{(k)}_\xi(r))^2\,dr \\
&\quad + \int_0^s p_x(r, \xi^*_{0,r}(x_0))g(r)\,dw_r + \sum_{k=1}^n \int_0^s p_r(\xi^*_{0,r}(x_0))g^{(k)}_\xi(r)\,dw^k_r \\
&\quad - \sum_{k=1}^n \int_0^s p_x(r, \xi^*_{0,r}(x_0))g(r)g^{(k)}_\xi(r)\,dr - \sum_{k=1}^n \int_0^s p_r(\xi^*_{0,r}(x_0))(g^{(k)}_\xi(r))^2\,dr \\
&= E[c_\xi(\xi^*_{0,T}(x_0))C_{0,T}] - \int_0^s p_r(\xi^*_{0,T}(x_0))f_\xi(r)\,dr \\
&\quad + \int_0^s p_x(r, \xi^*_{0,r}(x_0))g(r)\,dw_r - \sum_{k=1}^n \int_0^s p_x(r, \xi^*_{0,r}(x_0))g(r)g^{(k)}_\xi(r)\,dr,
\end{aligned}
$$

thus establishing the result.

This verifies by a simple, direct method the formula of Haussmann [12] without any requirement that the diffusion coefficient matrix gg^* is nonsingular. However we do not identify $p_s(x)$ with the gradient of the minimum cost process; this follows from arguments as in [12].

9. Conclusion. Using the theory of stochastic flows the effect of a perturbation of an optimal control is explicitly calculated in both the partially observed and stochastic open loop cases. The only difficulty is to justify the differentiation. The adjoint variable $p_s(x)$ is explicitly identified.

THEOREM 9.1. *If f is differentiable in the control variable u, and if the random variable $x = \xi^*_{0,s}(x_0)$ has a conditional density $q_s(x)$ under the measure P^*, then the inequality of Theorem* 5.1 *implies*

$$
\sum_{j=1}^k (u_j(x) - u^*_j(s)) \int_{R^d} \Gamma(s, x) \frac{\partial f}{\partial u_j}(s, x, u^*) q_s(x)\,dx \geqq 0.
$$

This is the result of Bensoussan's paper [1].

REFERENCES

[1] A. BENSOUSSAN, *Maximum principle and dynamic programming approaches of the optimal control of partially observed diffusions*, Stochastics, 9 (1983), pp. 169-222.

[2] J. N. BLAGOVESCENSKII AND M. I. FREIDLIN, *Some properties of diffusion processes depending on a parameter*, Dokl. Akad. Nauk SSSR, 138 (1961). (In Russian.) Soviet Math. Dokl., 2 (1961), pp. 633-636. (In English.)

[3] J. M. BISMUT, *Théorie probabiliste du contrôle des diffusions*, Memoirs American Mathematical Society 167, Providence, RI, 1976.

[4] ———, *An introductory approach to duality in optimal stochastic control*, SIAM Rev., 20 (1978), pp. 62-78.

[5] ———, *A generalized formula of Itô and some other properties of stochastic flows*, Z. Wahrsch. Verw. Gebiete, 55 (1981), pp. 331-350.

[6] ———, *Mécanique Aléatoire*, Lecture Notes in Mathematics 866, Springer-Verlag, Berlin, New York, 1981.

[7] J. M. BISMUT, *Mécanique Aléatoire*, In Ecole d'Eté de Probabilités de Saint-Flour X. Lecture Notes in Mathematics 929, Springer-Verlag, Berlin, New York, 1982, pp. 1-100.

[8] R. J. ELLIOTT, *The optimal control of a stochastic system*, SIAM J. Control Optim., 15 (1977), pp. 756-778.

[9] ———, *Filtering and Control for Point Process Observations*, Systems Science Center, University of Maryland, College Park, MD; Lecture Notes in Mathematics, Springer-Verlag, Berlin, New York, to appear.

[10] R. J. ELLIOTT AND M. KOHLMANN, *A short proof of a martingale representation result*, Statist. Probab. Lett., 6 (1988), pp. 327-329.

[11] U. G. HAUSSMANN, *General necessary conditions for optimal control of stochastic systems*, Math. Programming Stud., 6 (1976), pp. 30-48.

[12] ———, *On the adjoint process for optimal control of diffusion processes*, SIAM J. Control and Optim., 19 (1981), pp. 221-243, and 710.

[13] ———, *A Stochastic Minimum Principle for Optimal Control of Diffusions*, Pitman Research Notes in Mathematics 151, Longman, U.K., 1986.

[14] ———, *The maximum principle for optimal control of diffusions with partial information*, SIAM J. Control Optim., 25 (1987), pp. 341-361.

[15] N. V. KRYLOV, *Controlled Diffusion Processes*, Application of Mathematics Vol. 14, Springer-Verlag, Berlin, New York, 1980.

[16] H. KUNITA, *On the decomposition of solutions of stochastic differential equations*, Lecture Notes in Mathematics 851, Springer-Verlag, Berlin, New York, 1980, pp. 213-255.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1293-1318, November 1989

004

OPTIMAL AND APPROXIMATELY OPTIMAL CONTROL POLICIES FOR QUEUES IN HEAVY TRAFFIC*

HAROLD J. KUSHNER†‡ AND K. M. RAMACHANDRAN†§

Abstract. The "approximately" optimal control problem for tandem queueing or production networks (with local feedback allowed) under heavy traffic is treated. The buffers (scaled with traffic) are finite. The controls allow various inputs, connecting links, and the processors to be shut down or opened to manage the system. The service and arrival rates, as well as the routing probabilities, can also be controlled, and the system statistics can depend on the system state (scaled buffer occupancies). The associated costs involve holding costs, costs for shutting off/turning on the links or processors and the opportunity cost for lost production. It is shown that the (scaled) controlled system converges weakly (in an appropriate sense) to a controlled limit "reflected" diffusion. In the rescaled time, the actions of the controllers lead to multiple "simultaneous" impulses in the limit problem. Thus a nonstandard limit control problem is obtained, and the usual methods of weak convergence for systems under heavy traffic must be modified. Since it is usually not possible to obtain the optimal or nearly optimal controls for the physical process, it is of considerable interest to know whether an optimal or nearly optimal control for the limit process is also nearly optimal for the physical system with heavy traffic. This is shown to be true under reasonable conditions. Although the limit control problem is nonstandard and there is little available theory concerning it, acceptable numerical procedures are available.

Key words. weak convergence, queueing networks, production networks, heavy traffic approximations, controlled reflected diffusions, controlled queueing networks, approximately optimal stochastic controls, numerical methods for stochastic control

AMS(MOS) subject classifications. 93E20, 93E25, 90B22, 60F17, 60K25

1. Introduction.

We consider optimal and "nearly optimal" control problems for the open queueing networks in heavy traffic of the type dealt with in the fundamental papers of Reimann [1] and Harrison [2], [3]. Owing to the state and control dependence of the processes here, much of their methodology cannot be carried over. One of the main motivations behind the heavy traffic approximations [1]-[4] of queueing networks is the idea that the limit process is usually much easier to analyze than the actual physical process, and that it is much easier to find good control policies for the limit.

In [1], there are several interconnected service or processing stations, and at each there is an infinite buffer (ours is finite, but suitably scaled). At each there are possible arrivals from outside the network as well as arrivals routed from other service stations. Eventually with probability one (w.p.1) all customers leave the network. Under reasonable conditions on the interarrival and service times and with appropriate spatial and temporal normalizations, in the heavy traffic case the vector of the normalized queue lengths converges weakly to a reflected Brownian motion with constant drift and covariance parameters [1]. This will be generalized here in several directions, although we work with a somewhat simpler network structure.

Although it underlies much of the motivation for the limit theorems, there has been little work on the usefulness of the limit process for purposes of getting a good

* Received by the editors June 1, 1987; accepted for publication (in revised form) January 17, 1989.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

‡ The work of this author was supported in part by National Science Foundation contract ECS-8505674, Air Force Office of Scientific Research contract AFOSR-85-0315, and Office of Naval Research contract N00014-85-K-0607.

§ The work of this author was supported in part by Army Research Office contract DAAG29-84-K-0082 and Office of Naval Research contract N00014-85-K-0607. Present address, Department of Mathematics, University of South Florida, Tampa, Florida 33620.

or nearly optimum control for the physical process. Let ε index the traffic intensity. As $\varepsilon \to 0$, the "intensity" goes to one. For whatever cost criterion is used (this will be defined in later sections), let $V^\varepsilon(\pi)$ denote its value for the physical system when a policy π is used. Suppose that $\bar{\pi}^\varepsilon$ is an "adaptation" of the optimal or δ-optimal policy for the limit, applied to the physical process. (We will say more about such adaptations later.) For $\bar{\pi}^\varepsilon$ to be a "good" policy for the physical process we need at least that $V^\varepsilon(\bar{\pi}^\varepsilon) - \inf_\pi V^\varepsilon(\pi)$ be small for small ε, where the inf is over an appropriate set of policies for the physical process. This is the problem addressed here. In the course of the development, a number of interesting and nonclassical problems arise; for example, the appropriate "limit" control problem might involve multiple "simultaneous" impulses, and we must treat state-dependent service, arrival, and routing processes.

We choose a problem formulation that illustrates the main problems and allows the development of a method that applies to many other formulations. Our work differs from earlier work in several important respects. If the service or arrival rates can be controlled, then the limit process is no longer a reflected Brownian motion with constant coefficients. Owing to the control, there might be "travel" along the boundaries of the state space. Some control actions (e.g., on/off controls with associated impulsive costs) might yield a sequence of paths that do not converge in the Skorokhod topology, but there still is a meaningful sense in which the limit is a well-defined impulsively controlled process, perhaps with "multiple simultaneous impulses." The lumping together of all idle times as done in equation (3) of [1] in the $B_k(t)$ argument is a very slick idea, but it is inappropriate in our context owing to the state and control dependencies. We must show that the "limit" controls and other quantities are "admissible," or nonanticipative with respect to the limit Brownian motions. In fact, we combine the ideas of [1] with those of the martingale method and the weak convergence techniques of [5] and [6].

The present work is a continuation of the lines of development in [6]–[8] where approximations to other optimal control problems are dealt with.

In § 2, the basic system is described, the control problem is defined and the assumptions are stated. To avoid some quite complicated bookkeeping, we eventually specialize to the case where there are only two processors and feedback is only allowed from a processor to itself. The general results can be readily extended to problems where (except for the possibility of rerouting an output back to the input of the same processor), the flow is all "forward." In § 3, we discuss representations for the processes that facilitate the weak convergence analysis, and in § 4 we describe the proper "limit" control problem; i.e., the appropriate controlled reflected diffusion whose optimal (or δ-optimal) controls are to be used for the physical process.

Section 5 contains the basic weak convergence results, and we state and prove the results concerning the "almost optimality" of the δ-optimal (for small δ) controls for the limit process, when applied to the physical process. Some computational questions are discussed in § 6. Although the "limit" control problem is not always simple, effective and convenient numerical methods are available. Reference [10] contains an analysis of the dynamic programming equation for a one-dimensional version of the limit problem.

2. Problem description and assumptions. We start by describing a network with K service stations (processors), the ith referred to as P_i. Each processor services only one customer at a time, although batch or multiserver cases can all be handled. Shortly, we specialize to the case $K = 2$, but it is simpler to first use a unified terminology. We

retain the basic structure of [1], but use a discrete time parameter for notational simplicity. Each processor can be connected to an external input as well as receive (and deliver) outputs from (to) other processors.

Let $\{\alpha_n^{i,\varepsilon}\}$ denote the sequence of interarrival times of the customers coming to P_i from the exterior of the network, and let $\xi_n^{i,\varepsilon}$ denote the indicator of the event that there was an arrival from the exterior to P_i at time n. We use the convenient representation (as in [11]) where the processor keeps processing even if the queue is empty, with the "errors" generated by this convention accounted for by an added reflection term. With this convention in mind, let $\{\Delta_n^{i,\varepsilon}\}$ denote the sequence of service times for P_i, and $\psi_n^{i,\varepsilon}$ the indicator of the event that a service at P_i is completed at time n (whether or not there are actual "physical" customers in P_i at that time). As in [11], we suppose that if there is an arrival to P_i in the midst of a service interval when the queue at P_i is empty, then the actual service time for that customer is just the residual service time for the current service interval. Under the heavy traffic assumption, this does not affect the limit formulas. An outline of the proof is in the Appendix. Let $I_n^{ij,\varepsilon}$, $i=1,\cdots,K$, $j=0,\cdots,K$, denote the indicator function of the event that a completed service at P_i at time n is scheduled to be sent to P_j (or to the exterior, if $j=0$). We use $\{p_{ij}, i, j=1,\cdots,K\}$ to denote the probability that a completed service from P_i is to be routed to P_j, and write $p_{i0}=1-\sum_{j=1}^{K} p_{ij}$. The buffer size at P_i is $B_i/\sqrt{\varepsilon}$, for $B_i>0$.

We will impose the following assumptions: We work with impulsive controls only, although the results can be extended to the case where the service and interarrival "rates" are controlled continuously. The processor P_i can be shut off for a time, at a cost $k_i>0$, to be paid *at the moment of shut off*. The external inputs to P_i can be shut off for a time, at a cost $k_{0i}>0$, *to be paid at the moment of shut off*. If P_i communicates with P_j, in lieu of shutting off P_i, we can open or break the link connecting P_i to P_j. In that case the output of P_i destined for P_j will be shunted to the exterior and lost, or sold as a "partially completed" product. The cost for shutting off the link is $k_{ij}>0$, *to be paid at the moment of shut off*, and there will be an additional cost for the lost customers. This cost is $q_{ij}\sqrt{\varepsilon}$ per lost customer, $q_{ij}>0$. By convention, we allow all customers in P_i who have completed service there and are destined to return to P_i immediately to do so. If the buffer of P_i is full, then one or more inputs must be turned off, i.e., either the input links to P_i are shunted to the exterior, or the P_j connecting to P_i are shut off.

Let $P_n^{i,\varepsilon}$, $P_n^{0i,\varepsilon}$ and $P_n^{ji,\varepsilon}$, respectively, denote the indicators of the events that P_i is working at time n (i.e., processing or not shut off), the external input to P_i is not shut off at time n, and the link connecting P_j to P_i is open at time n, respectively. Let $N_n^{i,\varepsilon}$ (respectively, $\tilde{N}_n^{i,\varepsilon}$) denote the nth time that P_i is turned off (respectively, turned back on), and set $\tilde{N}_0^{i,\varepsilon}=0$. Let $N_n^{ij,\varepsilon}$ $(i=0,1,\cdots,K,\ j=1,\cdots,K)$ (respectively, $\tilde{N}_n^{ij,\varepsilon}$) denote the nth time that the link connecting P_i to P_j is shut off (turned back on, respectively). (If $i=0$, then it is for the link connecting the exterior to P_j.) Define $v_n^{i,\varepsilon}=\varepsilon N_n^{i,\varepsilon}$, $v_n^{ij,\varepsilon}=\varepsilon N_n^{ij,\varepsilon}$, and similarly define $\tilde{v}_n^{i,\varepsilon}$ and $\tilde{v}_n^{ij,\varepsilon}$.

Let $X_n^{i,\varepsilon}=\sqrt{\varepsilon}$ (number of customers in or waiting for service at P_i at time n) and set $X^{i,\varepsilon}(t)=X_{t/\varepsilon}^{i,\varepsilon}$. This is the quantity of interest in the desired interpolated time and amplitude scale. Then, in this interpolated scale, $[v_n^{i,\varepsilon}, \tilde{v}_n^{i,\varepsilon})$, $n\geq 1$, etc., are the intervals of closure of P_i, etc. When ratios t/ε are used as indices, we use the integral part. Until §§ 5 and 6, and for notational convenience, we always assume that all processors and links are working at $t=0$. Thus $\tilde{v}_0^{\alpha}\equiv 0$ and $\tilde{v}_n^{\alpha,\varepsilon}>v_n^{\alpha,\varepsilon}$ for $n>0$.

To keep track of the flows in the system in a way that allows a convenient development of the limit theorems, we need to separate the corrections to the flows due to empty queues and to the flow components due to the control actions. This is

why (2.1)-(2.4) are introduced. *Throughout the paper, ε-superscripts will be omitted in the terms in sums or integrals.* Define

$$Y^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^i I_n^{ij} P_n^i I_{\{X_n^i = 0\}}, \tag{2.1a}$$

$$U^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^i I_n^{ij} (1 - P_n^i), \qquad i \neq 0, \quad j \neq i, \tag{2.1b}$$

$$U^{0i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \xi_n^i (1 - P_n^{0i}), \qquad i \neq 0, \tag{2.2a}$$

$$U_c^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^j I_n^{ji} (1 - P_n^{ji} P_n^j), \qquad j \neq i, \quad j \neq 0, \tag{2.2b}$$

$$Y_c^{ji,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^j I_n^{ji} P_n^j P_n^{ji} I_{\{X_n^j = 0\}}, \qquad j \neq i, \quad j \neq 0, \tag{2.2c}$$

$$A^{i,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \xi_n^i, \quad D^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^i I_n^{ij}, \quad i \neq 0, \tag{2.3}$$

$$Z^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{n=1}^{t/\varepsilon} \psi_n^i I_n^{ij} (1 - P_n^{ij}) P_n^i I_{\{X_n^i \neq 0\}}, \qquad i \neq j, \quad i, j \neq 0. \tag{2.4}$$

With the definitions (2.1)-(2.3), we can write

$$\begin{aligned} X^{i,\varepsilon}(t) = {} & A^{i,\varepsilon}(t) + \sum_{j \neq i} D^{ji,\varepsilon}(t) - \sum_{j \neq i} D^{ij,\varepsilon}(t) + \sum_{j \neq i} Y^{ij,\varepsilon}(t) \\ & - \sum_{j \neq i} Y_c^{ji,\varepsilon}(t) - U^{0i,\varepsilon}(t) + \sum_{j \neq i} U^{ij,\varepsilon}(t) - \sum_{j \neq i} U_c^{ji,\varepsilon}(t). \end{aligned} \tag{2.5}$$

The first term in (2.5) respresents the potential external arrivals to P_i, the second represents potential arrivals from other P_j, $j \neq i$, all neglecting the effects of controls or empty queues. The third term represents potential departures from P_i, again neglecting the effects of controls or empty queues. The other terms correct for these omissions. The $Y^{ij,\varepsilon}(\cdot)$ corrects for departures from P_i when P_i is working and its queue is empty, and the $Y_c^{ji,\varepsilon}(\cdot)$ corrects for arrivals to P_i from P_j when the buffer of P_j is empty and neither P_j nor the link from P_j to P_i is shut off. The $U^{0i,\varepsilon}(\cdot)$ corrects for the stopped external arrivals, when the input to P_i from the exterior is shut off. The $U^{ij,\varepsilon}(\cdot)$ corrects for the stopped departures from P_i when P_i is closed, and the $U_c^{ji,\varepsilon}(\cdot)$ corrects for the stopped arrivals from P_j to P_i when either P_j is not working or the link from P_j to P_i is shut off (i.e., shunted to the exterior).

The $Z^{ij,\varepsilon}(\cdot)$ represents the lost output when the link from P_i to P_j is shunted to the exterior. There can only be lost output at time n if $X_n^{i,\varepsilon} > 0$ and $P_n^{i,\varepsilon} = 1$ and $P_n^{ij,\varepsilon} = 0$. Write $X^\varepsilon = (X^{1,\varepsilon}, \cdots, X^{K,\varepsilon})$ and let π^ε or π denote control policies (i.e., rules for determining the $v^{i,\varepsilon}$, $\tilde{v}^{i,\varepsilon}$, $v^{ij,\varepsilon}$, $\tilde{v}^{ij,\varepsilon}$), and let E_x^π denote the expectation, given policy π and initial condition $X_0^\varepsilon = x$. Let P_t denote the *vector* of *indicator functions* $\{P_t^\alpha\}$ of the processors and links at time t. (We set $P_0^\alpha = 1$ until § 5). Then, for a bounded and continuous $k(\cdot)$ and $\beta > 0$, our cost will be of the discounted form:

$$\begin{aligned} V^\varepsilon(\pi, x, P) = {} & E_x^\pi \int_0^\infty e^{-\beta t} k(X^\varepsilon(t))\, dt + E_x^\pi \sum_{i=1}^K k_i \sum_n e^{-\beta v_n^{i,\varepsilon}} + E_x^\pi \sum_{i=0}^K \sum_{j=1}^K k_{ij} \sum_n e^{-\beta v_n^{ij,\varepsilon}} \\ & + E_x^\pi \int_0^\infty e^{-\beta t} \left[\sum_i q_{0i}\, dU^{0i,\varepsilon}(t) + \sum_{i,j=1}^K q_{ij}\, dZ^{ij,\varepsilon}(t) \right]. \end{aligned} \tag{2.6}$$

We now specialize to the case of Fig. 2.1, since it is very awkward to keep track of the effects of the controls in a network with general feedback allowed. With mainly notational changes, the case dealt with here can be extended to the general feedforward case.

Refer to Fig. 2.1, and assume (A2.1). The first part of this assumption says that if a queue is empty, then we will not continue to "starve" it—but will turn on all the inputs. The assumption seems to be quite unrestrictive, and it does simplify the bookkeeping quite a bit.

(A2.1) If $X_n^{2,\varepsilon}=0$, then all inputs to P_2 are open; i.e., $P_n^{1,\varepsilon}=P_n^{12,\varepsilon}=P_n^{02,\varepsilon}=1$. If $X_n^{1,\varepsilon}=0$, then the input to P_1 is open (i.e., $P_n^{01,\varepsilon}=1$). If some $X_n^{i,\varepsilon}=B_i$, then all inputs to P_i are closed.

For the system of Fig. 2.1, and under (A2.1), we have that (2.1)-(2.5) take the forms (2.7)-(2.9). Here, $P_n^{2,\varepsilon}\equiv 1$, since there is never a need to shut off P_2:

$$
\begin{aligned}
Y_c^{12,\varepsilon}(t) &= \sqrt{\varepsilon}\sum_1^{t/\varepsilon}\psi_n^1 I_n^{12} P_n^1 P_n^{12} I_{\{X_n^1=0\}},\\
Y^{20,\varepsilon}(t) &= \sqrt{\varepsilon}\sum_1^{t/\varepsilon}\psi_n^2 I_n^{20} I_{\{X_n^2=0\}},\\
Z^{12,\varepsilon}(t) &= \sqrt{\varepsilon}\sum_1^{t/\varepsilon}\psi_n^1 I_n^{12}(1-P_n^{12})P_n^1 I_{\{X_n^1\neq 0\}}\\
&= U_c^{12,\varepsilon}(t)-U^{12,\varepsilon}(t)-\sum_1^{\infty}\int_{v_n^{12}\cap t}^{\tilde v_n^{12}\cap t} dY^{12,\varepsilon}(s).
\end{aligned}
\tag{2.7}
$$

The $Y^{12,\varepsilon}(\cdot)$ will converge to a continuous function and $\tilde v_n^{12,\varepsilon}-v_n^{12,\varepsilon}\overset{\varepsilon}{\to}0$. Thus the last term on the right of the last equation will disappear in the limit. Define $U^{1,\varepsilon}(\cdot)=U^{10,\varepsilon}(\cdot)+U^{12,\varepsilon}(\cdot)$. Then

$$X^{1,\varepsilon}(t)=A^{1,\varepsilon}(t)-D^{10,\varepsilon}(t)-D^{12,\varepsilon}(t)+Y^{10,\varepsilon}(t)+Y^{12,\varepsilon}(t)-U^{01,\varepsilon}(t)+U^{1,\varepsilon}(t), \tag{2.8a}$$

$$X^{2,\varepsilon}(t)=A^{2,\varepsilon}(t)-D^{20,\varepsilon}(t)+D^{12,\varepsilon}(t)+Y^{20,\varepsilon}(t)-Y_c^{12,\varepsilon}(t)-U^{02,\varepsilon}(t)-U_c^{12,\varepsilon}(t), \tag{2.8b}$$

$$
\begin{aligned}
V^{\varepsilon}(\pi,x,P) &= E_x^{\pi}\int_0^{\infty} e^{-\beta t}k(X^{\varepsilon}(t))\,dt + k_1 E_x^{\pi}\sum_n e^{-\beta v_n^{1,\varepsilon}} + \sum_{i=1}^{2} k_{0i}E_x^{\pi}\sum_n e^{-\beta v_n^{0i,\varepsilon}}\\
&\quad + k_{12}E_x^{\pi}\sum_n e^{-\beta v_n^{12,\varepsilon}} + E_x^{\pi}\int_0^{\infty} e^{-\beta t}\left[\sum_{i=1}^{2} q_{0i}\,dU^{0i,\varepsilon}(t) + q_{12}\,dZ^{12,\varepsilon}(t)\right].
\end{aligned}
\tag{2.9}
$$

We now give some more definitions and state the heavy traffic assumptions. It will sometimes be convenient to write the multiple sequence $v^{\varepsilon}\equiv\{v_n^{i,\varepsilon}, \tilde v_n^{i,\varepsilon}, v_n^{ij,\varepsilon}, \tilde v_n^{ij,\varepsilon}\}$

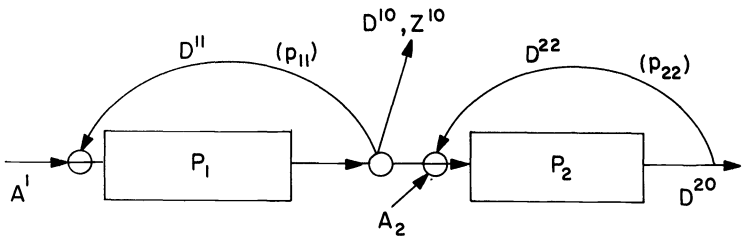

FIG. 2.1. *The system configuration.*

as a single sequence. Let $\{\tau_n^\varepsilon\}$ denote the sequence of event *times* indicated by all the elements of v^ε in nonincreasing order, but without respect to which events they indicate, or whether they indicate multiple events. Define $R_n^\varepsilon = (R_n^{1,\varepsilon}, R_n^{01,\varepsilon}, R_n^{02,\varepsilon}, R_n^{12,\varepsilon})$, where $R_n^{\alpha,\varepsilon} = 1, -1$, or 0 depending on whether or not the "control" with the same superscript was opened (turned on), closed (turned off) or left unchanged at τ_n^ε. *From* $\{R_n^\varepsilon, \tau_n^\varepsilon, \delta U_n^\varepsilon\}$, *we can recover all the control actions and their times and values.*

Define

$$\delta U_n^\varepsilon = (U^{1,\varepsilon}(\tau_{n+1}^\varepsilon) - U^{1\varepsilon}(\tau_n^\varepsilon),\ U^{01,\varepsilon}(\tau_{n+1}^\varepsilon) - U^{01,\varepsilon}(\tau_n^\varepsilon),$$
$$U^{02,\varepsilon}(\tau_{n+1}^\varepsilon) - U^{02,\varepsilon}(\tau_n^\varepsilon),\ U_c^{12,\varepsilon}(\tau_{n+1}^\varepsilon) - U_c^{12,\varepsilon}(\tau_n^\varepsilon)).$$

If τ_n^ε is not defined for some n, ω, then let it equal to infinity there, and define the associated R_n^ε and δU_n^ε arbitrarily.

Let $S_{a,n}^{i,\varepsilon} = \sum_{j=1}^n \alpha_j^{i,\varepsilon}$, $S_{d,n}^{i,\varepsilon} = \sum_{j=1}^n \Delta_j^{i,\varepsilon}$. Let $E_{a,n}^{i,\varepsilon}$ denote the expectation given the arrival, departure, and control intervals and actions that ended by real time $S_{a,n}^{i,\varepsilon}$, as well as the lengths of all other arrival and service intervals (other than $\alpha_{n+1}^{i,\varepsilon}$) that started by, but might not have been completed by, time $S_{a,n}^{i,\varepsilon}$. Analogously, $E_{d,n}^{i,\varepsilon}$ denotes the expectation given the arrival, departure, and control intervals and actions that ended by real time $S_{d,n}^{i,\varepsilon}$, as well as the lengths of all other arrival and service intervals (other than $\Delta_{n+1}^{i,\varepsilon}$) that started by $S_{d,n}^{i,\varepsilon}$. Define the conditional variances $\mathrm{var}_{a,n}^{i,\varepsilon}$, $\mathrm{var}_{d,n}^{i,\varepsilon}$ analogously.

Define

$$E_{a,n}^{i,\varepsilon}\alpha_{n+1}^{i,\varepsilon} = \bar{\alpha}_{n+1}^{i,\varepsilon}, \qquad \mathrm{var}_{a,n}^{i,\varepsilon}\ \alpha_{n+1}^{i,\varepsilon} = (\sigma_{a,n+1}^{i,\varepsilon})^2,$$
$$E_{d,n}^{i,\varepsilon}\Delta_{n+1}^{i,\varepsilon} = \bar{\Delta}_{n+1}^{i,\varepsilon}, \qquad \mathrm{var}_{d,n}^{i,\varepsilon}\ \Delta_{n+1}^{i,\varepsilon} = (\sigma_{d,n+1}^{i,\varepsilon})^2.$$

Henceforth when we say that P_i, P_{0i}, *or* P_{12}, *respectively, is open* (*closed*) *at time n, we mean that processor i is working, the link from the exterior to* P_i *is open or* (*respectively*), *the link from* P_1 *to* P_2 *is open for traffic.*

We will use the following assumption.

(A2.2) There are positive numbers g_{ai} and g_{di} and bounded continuous functions $a^i(\cdot)$ and $d^i(\cdot)$ such that

$$[\bar{\alpha}_{n+1}^{i,\varepsilon}]^{-1} = g_{ai} + \sqrt{\varepsilon}\ a_{in} + o(\sqrt{\varepsilon}), \qquad [\bar{\Delta}_{n+1}^{i,\varepsilon}]^{-1} = g_{di} + \sqrt{\varepsilon}\ d_{in} + o(\sqrt{\varepsilon}),$$

where $a_{in} = a^i(X_{S_{a,n}^{i,\varepsilon}}^\varepsilon)$ and $d_{in} = d^i(X_{S_{d,n}^{i,\varepsilon}}^\varepsilon)$.

Comment on (A2.2). We allow the (marginal) external interarrival intervals and the service intervals to depend on the system state. The argument $X_{S_{a,n}^{i,\varepsilon}}^\varepsilon$ (for example) is the proper one, since $S_{a,n}^{i,\varepsilon}$ is the moment of arrival to P_i of the $(n+1)$st customer from the outside, and $X_{S_{a,n}^{i,\varepsilon}}^\varepsilon$ is the system state at that time. At some expense in details, we could let the marginal mean rates $a^i(\cdot)$ and $d^i(\cdot)$ be controlled. We would then use the forms $a^i(X_{S_{a,n}^{i,\varepsilon}}^i, r_{S_{a,n}^{i,\varepsilon}}^{i,\varepsilon})$ etc., where the r_β^α represents controls. The condition (A2.2) together with (A2.4) imply that the total load put on processor P_i is very close to its processing capacity. In other words, the idle time is negligible in the sense that it converges to zero as $\varepsilon \downarrow 0$.

(A2.3) The set $\{|\alpha_n^{i,\varepsilon}|^2, |\Delta_n^{i,\varepsilon}|^2, i, n < \infty$, small ε, all control actions$\}$ is uniformly integrable.

(A2.4) (Heavy traffic assumption)

$$g_{a1} = (1 - p_{11})g_{d1}, \qquad [p_{12}g_{d1} + g_{a2}]/(1 - p_{22}) = g_{d2}.$$

Assumption (A2.4) is also what we would get from Reimann's [1] formulas for the case of Fig. 2.1.

(A2.5) The routing variables $\{I_k^{ij,\varepsilon}, i, j, k\}$ are mutually independent and independent of the $\{\alpha_k^{i,\varepsilon}, \Delta_k^{i,\varepsilon}\}$ and $P\{I_k^{ij,\varepsilon}=1\}=p_{ij}$.

(A2.6) There are continuous functions $\sigma_{ai}(\cdot)$, $\sigma_{di}(\cdot)$ such that

$$\sigma_{a,n+1}^{i,\varepsilon}=\sigma_{a,i}(X^{\varepsilon}_{S_{a,n}^{i,\varepsilon}})+\delta'_{\varepsilon}, \qquad \sigma_{d,n+1}^{i,\varepsilon}=\sigma_{d,i}(X^{\varepsilon}_{S_{d,n}^{i,\varepsilon}})+\delta''_{\varepsilon}$$

where $\delta_{\varepsilon}^{\alpha} \overset{\varepsilon}{\to} 0$, uniformly in all other variables.

Comment on (A2.5) *and* (A2.6). We allow the conditional variance to depend on the state here, just to show the possibilities. The sequence of interarrival times or service intervals can be correlated (in ways other than via the "state" dependence used here). This would involve a more complex method for obtaining the weak convergence. The perturbed test function methods of [5] can be used, but the additional notational burden hardly seems worth it now.

3. A convenient representation for $X^{\varepsilon}(\cdot)$. In this section, we center and rewrite the terms of (2.8) to facilitate the weak convergence analysis in § 5. We will do three things. First, the A and D processes will be centered, the centering terms simplified, and the centered processes written as a rescaling of simpler processes. This is similar to the procedure of [1]. Then we will represent the $Y^{ij,\varepsilon}$ and $U^{ij,\varepsilon}$ in terms of simpler processes $Y^{i,\varepsilon}$ and $U^{i,\varepsilon}$ plus a term that will go to zero as $\varepsilon\to 0$. Finally, we will represent $Y^{i,\varepsilon}$ and $X^{i,\varepsilon}$ as continuous functions of the "other" data.

Centering of the arrival and departure processes. Define $\bar{S}_a^{i,\varepsilon}(t)$ (and analogously $\bar{S}_d^{i,\varepsilon}(t)$) to be the *inverse* of the interpolated arrival time function $\varepsilon S_{a,t/\varepsilon}^{i,\varepsilon}$. More precisely, define

$$\bar{S}_a^{i,\varepsilon}(t)=\max\{\varepsilon k\colon \varepsilon S_{a,k}^{i,\varepsilon}\leqq t\}.$$

Define the centered processes

$$\begin{aligned}\tilde{A}_0^{i,\varepsilon}(t)&=\sqrt{\varepsilon}\sum_{k=1}^{t/\varepsilon}\sum_{l=S_{a,k}^{i,\varepsilon}}^{S_{a,k+1}^{i,\varepsilon}-1}\left[\xi_l^i-\frac{1}{\bar{\alpha}_k^i}\right]\\&=\sqrt{\varepsilon}\sum_{1}^{t/\varepsilon}\left(1-\frac{\alpha_k^i}{\bar{\alpha}_k^i}\right),\\ \tilde{D}_0^{ij,\varepsilon}(t)&=\sqrt{\varepsilon}\sum_{k=1}^{t/\varepsilon}\sum_{l=S_{d,k}^{i,\varepsilon}}^{S_{d,k+1}^{i,\varepsilon}-1}\left[\psi_l^i I_l^{ij}-\frac{p_{ij}}{\bar{\Delta}_k^i}\right]\\&=\sqrt{\varepsilon}\sum_{k=1}^{t/\varepsilon}\left[I^{ij}_{S_{d,k}^i}-p_{ij}\frac{\Delta_k^i}{\bar{\Delta}_k^i}\right].\end{aligned} \tag{3.1}$$

The second equality in the first definition follows from the fact that $\xi_l^{i,\varepsilon}=1$ only at the left endpoint in the interval $[S_{a,k}^{i,\varepsilon}, S_{a,k+1}^{i,\varepsilon})$ and the length of the interval is $\alpha_k^{i,\varepsilon}$.

Owing to the independence assumptions in (A2.5), we can replace the $I^{ij}_{S_{d,k}^i}$ by I_k^{ij}. We can write the $A^{i,\varepsilon}(\cdot)$ of (2.3) in the form

$$\begin{aligned}A^{i,\varepsilon}(t)&=\sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}_a^{i,\varepsilon}(t)}\sum_{l=S_{a,k}^{i,\varepsilon}}^{S_{a,k+1}^{i,\varepsilon}-1}\left[\xi_l^i-\frac{1}{\bar{\alpha}_k^i}\right]+\sqrt{\varepsilon}\sum_{k=1}^{\varepsilon^{-1}\bar{S}_a^{i,\varepsilon}(t)}\frac{\alpha_k^i}{\bar{\alpha}_k^i}\\&\equiv\tilde{A}_0^{i,\varepsilon}(\bar{S}_a^{i,\varepsilon}(t))+\tilde{B}_a^{i,\varepsilon}(t)\equiv\tilde{A}^{i,\varepsilon}(t)+\tilde{B}_a^{i,\varepsilon}(t).\end{aligned} \tag{3.2}$$

Doing the same thing for the $D^{ij,\varepsilon}(\cdot)$, we have

$$D^{ij,\varepsilon}(t) = \tilde{D}_0^{ij,\varepsilon}(\bar{S}_d^{i,\varepsilon}(t)) + \tilde{B}_d^{ij,\varepsilon}(t) \equiv \tilde{D}^{ij,\varepsilon}(t) + \tilde{B}_d^{ij,\varepsilon}(t) \tag{3.3}$$

where

$$\tilde{B}_d^{ij,\varepsilon}(t) = \sqrt{\varepsilon} \sum_{k=1}^{\varepsilon^{-1}\bar{S}_d^{i,\varepsilon}(t)} \frac{\Delta_k^i}{\bar{\Delta}_k^i} p_{ij}. \tag{3.4}$$

For purposes of calculation below, write

$$\tilde{D}_0^{10,\varepsilon}(t) + \tilde{D}_0^{12,\varepsilon}(t) = \sqrt{\varepsilon} \sum_1^{t/\varepsilon} \left[(1 - I_k^{11}) - (1 - p_{11}) \frac{\Delta_k^i}{\bar{\Delta}_k^i} \right].$$

We now cancel the "principal parts" of the $\tilde{B}_\alpha^{i,\varepsilon}$ terms. By taking the terms in the order in which they would appear in the centering of the first three terms of (2.8a) and using the expansion in (A2.2), we write

$$\begin{aligned} \tilde{B}_a^{1,\varepsilon}(t) - (\tilde{B}_d^{10,\varepsilon}(t) + \tilde{B}_d^{12,\varepsilon}(t)) &= \sqrt{\varepsilon} \sum_{k=1}^{\varepsilon^{-1}\bar{S}_a^{1,\varepsilon}(t)} \alpha_k^1 [g_{a1} + \sqrt{\varepsilon}\, a_{1k} + o(\sqrt{\varepsilon})] \\ &\quad - \sqrt{\varepsilon} \sum_{k=1}^{\varepsilon^{-1}\bar{S}_d^{1,\varepsilon}(t)} \Delta_k^1 [g_{d1} + \sqrt{\varepsilon}\, d_{1k} + o(\sqrt{\varepsilon})](1 - p_{11}). \end{aligned} \tag{3.5}$$

Since $\sum_{\varepsilon k=1}^{\bar{S}_a^{1,\varepsilon}(t)} \alpha_k^1 = 1/\varepsilon \pmod{O(1)}$, the principal term of the first sum is $g_{a1}t/\sqrt{\varepsilon}$ $(\text{mod } O(\sqrt{\varepsilon}))$, and of the second is $(1-p_{11})g_{d1}t/\sqrt{\varepsilon}$ $(\text{mod } O(\sqrt{\varepsilon}))$. These cancel by (A2.4). By using the definitions of α_k^1 and a_{1k}, we can write the sum of the middle terms in the first sum of (3.5) as

$$\varepsilon \sum_1^{t/\varepsilon} a^1(X_k) + \delta_\varepsilon^0(t)$$

(where all $\delta_\varepsilon^i(\cdot)$ here and below go to zero uniformly on bounded time intervals as $\varepsilon \to 0$), and similarly for the analogous terms in the second sum.

With the above cancellations and the last representation, we can rewrite (3.5) as

$$\begin{aligned} \varepsilon \sum_1^{t/\varepsilon} [a^1(X_k) - (1-p_{11})d^1(X_k)] + \delta_\varepsilon^1(t) &\equiv \int_0^t b^1(X^\varepsilon(s))\, ds + \delta_\varepsilon^1(t) \\ &\equiv B^{1,\varepsilon}(t) + \delta_\varepsilon^1(t). \end{aligned} \tag{3.6}$$

Repeating the procedure for the "biases" arising from (2.8b), we get

$$\begin{aligned} \tilde{B}_a^{2,\varepsilon}(t) - \tilde{B}_d^{20,\varepsilon}(t) + \tilde{B}_d^{12,\varepsilon}(t) &\equiv B_2^{2,\varepsilon}(t) + \delta_\varepsilon^2(t) = \int_0^t b^2(X^\varepsilon(s))\, ds + \delta_\varepsilon^2(t) \\ &= \varepsilon \sum_1^{t/\varepsilon} [a^2(X_k) - (1-p_{22})d^2(X_k) + p_{12}d^1(X_k)] + \delta_\varepsilon^2(t) \\ &= \int_0^t b^2(X^\varepsilon(s))\, ds + \delta_\varepsilon^2(t) = B^{2,\varepsilon}(t) + \delta_\varepsilon^2(t). \end{aligned} \tag{3.7}$$

A representation for $U^{ij,\varepsilon}$, $Y^{ij,\varepsilon}$. Define the processes (with $P_n^{2,\varepsilon} \equiv 1$)

$$Y^{1,\varepsilon}(\cdot) = Y^{10,\varepsilon}(\cdot) + Y^{12,\varepsilon}(\cdot), \qquad Y^{2,\varepsilon}(\cdot) = Y^{20,\varepsilon}(\cdot). \tag{3.8}$$

We can also write

$$U^{12,\varepsilon}(t) = \sum_{n=1}^{\infty} \int_{v_n^{1,\varepsilon} \cap t}^{\tilde{v}_n^{1,\varepsilon} \cap t} dU_c^{12,\varepsilon}(s). \tag{3.9a}$$

It will turn out (see § 5) that the limits in (3.9b) hold:

$$
\begin{aligned}
&U^{1j,\varepsilon}(\cdot) - U^{1,\varepsilon}(\cdot)p_{1j}/(1-p_{11}) \Rightarrow 0, \qquad j = 0, 2,\\
&U_c^{12,\varepsilon}(t) - \sqrt{\varepsilon}\, p_{12} \sum_0^{t/\varepsilon} \psi_n^1 (1 - P_n^1 P_n^{12}) \Rightarrow 0,\\
&Z^{12,\varepsilon}(\cdot) - [U_c^{12,\varepsilon}(\cdot) - U^{12,\varepsilon}(\cdot)] \Rightarrow 0,\\
&\tilde{v}_n^{\alpha,\varepsilon} - v_n^{\alpha,\varepsilon} \to 0 \quad \text{for all } \alpha, n,\\
&Y^{12,\varepsilon}(\cdot) - \frac{p_{12}}{p_{10}+p_{12}}\, Y^{1,\varepsilon}(\cdot) \Rightarrow 0,\\
&Y_c^{12,\varepsilon}(\cdot) - Y^{12,\varepsilon}(\cdot) \Rightarrow 0.
\end{aligned}
\tag{3.9b}
$$

To prepare for the utilization of these convergences and simplifications, rewrite (2.8) as (3.10), where the $\rho^{i,\varepsilon}(\cdot)$ are linear combinations of the $\delta_\varepsilon(\cdot)$ in (3.6) and (3.7) and $\hat{\rho}^{2,\varepsilon}(\cdot) = \rho^{2,\varepsilon}(\cdot) + (Y_c^{12,\varepsilon}(\cdot) - p_{12}Y^{1,\varepsilon}(\cdot)/(p_{10}+p_{12}))$, and the $W^{i,\varepsilon}(\cdot)$, $i = 1, 2$, are defined to be the sum of the first three terms in the middle part of (3.10a) and (3.10b), respectively:

$$
\begin{aligned}
X^{1,\varepsilon}(t) &= \tilde{A}^{1,\varepsilon}(t) - (\tilde{D}^{10,\varepsilon}(t) + \tilde{D}^{12,\varepsilon}(t)) + B^{1,\varepsilon}(t)\\
&\quad + (Y^{10,\varepsilon}(t) + Y^{12,\varepsilon}(t)) - U^{01,\varepsilon}(t) + U^{1,\varepsilon}(t) + \rho^{1,\varepsilon}(t)\\
&= W^{1,\varepsilon}(t) + B^{1,\varepsilon}(t) + Y^{1,\varepsilon}(t) - U^{0,1\varepsilon}(t) + U^{1,\varepsilon}(t) + \rho^{1,\varepsilon}(t),
\end{aligned}
\tag{3.10a}
$$

$$
\begin{aligned}
X^{2,\varepsilon}(t) &= \tilde{A}^{2,\varepsilon}(t) - \tilde{D}^{20,\varepsilon}(t) + \tilde{D}^{12,\varepsilon}(t) + B^{2,\varepsilon}(t)\\
&\quad + Y^{20,\varepsilon}(t) - Y_c^{12,\varepsilon}(t) - U^{02,\varepsilon}(t) - U_c^{12,\varepsilon}(t) + \rho^{2,\varepsilon}(t)\\
&= W^{2,\varepsilon}(t) + B^{2,\varepsilon}(t) + Y^{2,\varepsilon}(t) - p_{12}Y^{1,\varepsilon}(t)/(p_{12}+p_{10})\\
&\quad - U^{02,\varepsilon}(t) - U_c^{12,\varepsilon}(t) + \hat{\rho}^{2,\varepsilon}(t).
\end{aligned}
\tag{3.10b}
$$

We also write

$$
V^\varepsilon(\pi, x) = [(2.9) \text{ with } Z^{12,\varepsilon}(\cdot) = U_c^{12,\varepsilon}(\cdot) - U^{12,\varepsilon}(\cdot) + \hat{\rho}^{3,\varepsilon}(\cdot)]
\tag{3.11}
$$

where $\hat{\rho}^{3,\varepsilon}(\cdot)$ is an "error" term. We have $\sup_{t \leqq T} |\rho^{i,\varepsilon}(t)| \to 0$ in distribution as $\varepsilon \to 0$, for each $T < \infty$. Also, it will be shown in § 5 that, for any sequence of controls π^ε with $\sup V^\varepsilon(\pi^\varepsilon, x) < \infty$, $\sup_{t \leqq T} |\hat{\rho}^{i,\varepsilon}(t)| \to 0$ in distribution for any $T < \infty$.

Owing to the impulsive nature of the "control" part of the cost (2.9), on any bounded time interval there are only a finite number (w.p.1) of subintervals on which the controls are active. By the definitions, the reflection terms $Y^{ij,\varepsilon}(\cdot)$ cannot increase on these "control intervals." In particular, $Y^{1,\varepsilon}(\cdot)$ can only increase when both P_{01} and P_1 are on (recall that P_{01} must be on when $X^1 = 0$). Also, $Y^{2,\varepsilon}(\cdot)$ can increase only when all of P_1, P_{12}, and P_{02} are on (by (A2.1), if $X_n^{2,\varepsilon} = 0$, then all inputs must be turned on). Because of this and the feedforward nature of the problem, the simplest form of the reflection principle can be used to obtain the "reflection" terms as continuous functions of the other "noncontrol" data, simply by working with the appropriate "noncontrol" time segments, and we now formalize this.

The following result is well known and is a special case of the cited results in [1] and [12].

LEMMA 3.1. *Let $z(\cdot)$ be in $D[0, \infty)$, the space of real-valued functions with left-hand limits and that are right continuous, and with the* sup *norm topology. There is a unique $y(\cdot)$ in $D[0, \infty)$ such that $x(\cdot) = z(\cdot) + y(\cdot)$ and $x(t) \geqq 0$, $y(0) = 0$, and $y(\cdot)$ is nondecreasing and increases only when $x(\cdot) = 0$. In particular, $y(t) = -\min\{0, \inf_{s \leqq t} z(s)\}$. The map $z \to y$ is continuous in the* sup *norm on each finite interval* $[0, T]$.

Remark. In all cases below, the functions that replace z, x, and y will be the obvious terms from (3.10). We will want a representation of $y^\varepsilon(\cdot)$ as a continuous function of the other terms on the right-hand side of (3.10) to simplify the weak convergence proof.

Let $J_n^{1,\varepsilon} = [\mu_n^{1,\varepsilon}, \tilde{\mu}_n^{1,\varepsilon})$ denote the sequence of successive intervals (of interpolated time) such that $P_k^{1,\varepsilon} = P_k^{01,\varepsilon} = 1$ for $\varepsilon k \in J_n^{1,\varepsilon}$, and let $J_n^{2,\varepsilon} = [\mu_n^{2,\varepsilon}, \tilde{\mu}_n^{2,\varepsilon})$ denote the successive intervals such that $P_k^{1,\varepsilon} = P_k^{12,\varepsilon} = P_k^{02,\varepsilon} = 1$ for $\varepsilon k \in J_n^{2,\varepsilon}$. The $Y^{i,\varepsilon}(\cdot)$ can increase only on the $J_n^{i,\varepsilon}$.

We apply the representation in Lemma 3.1 to get an alternative representation of the increments of $Y^{i,\varepsilon}(\cdot)$ on the time segments between the control actions. For any function $f(\cdot)$ in $D[0,\infty)$, define the function $\delta_{in}^\varepsilon f(\cdot)$ in $D[0,\infty)$ by $\delta_{in}^\varepsilon f(\cdot) = f((\mu_n^{i,\varepsilon}+\cdot) \cap \tilde{\mu}_n^{i,\varepsilon}) - f(\mu_n^{i,\varepsilon})$. The function $\delta_{in}^\varepsilon f(\cdot)$ is just the segment of the function $f(\cdot)$, stopped at $\tilde{\mu}_n^{i,\varepsilon}$, shifted left by $\mu_n^{i,\varepsilon}$ and centered by subtracting the new "initial value" $f(\mu_n^{i,\varepsilon})$. We now apply Lemma 3.1 on the intervals $J_n^{1,\varepsilon}$ and $J_n^{2,\varepsilon}$ in turn.

By Lemma 3.1, for $t \geqq 0$ we have

$$\delta_{1n}^\varepsilon Y^{1,\varepsilon}(t) = -\min\left[\inf_{s \leqq t}(X^{1,\varepsilon}(\mu_n^{1,\varepsilon}) + \delta_{1n}^\varepsilon W^{1,\varepsilon}(s) + \delta_{1n}^\varepsilon B^{1,\varepsilon}(s) + \delta_{1n}^\varepsilon \rho^{1,\varepsilon}(s)), 0\right],$$

$$\delta_{2n}^\varepsilon Y^{2,\varepsilon}(t) = -\min\left[\inf_{s \leqq t}\left(X^{2,\varepsilon}(\mu_n^{2,\varepsilon}) + \delta_{2n}^\varepsilon W^{2,\varepsilon}(s) + \delta_{2\varepsilon}^\varepsilon B^{2,\varepsilon}(s) + \frac{p_{12}}{p_{12}+p_{10}}\delta_{2\varepsilon}^\varepsilon Y^{1,\varepsilon}(s) + \delta_{2n}^\varepsilon \hat{\rho}^{2,\varepsilon}(s)\right), 0\right]. \tag{3.12}$$

Also, we have

$$Y^{i,\varepsilon}(t) = \sum_{n:\, \mu_n^{i,\varepsilon} \leqq t} \delta_{in}^\varepsilon Y^{i,\varepsilon}(t - \mu_n^{i,\varepsilon}).$$

We use the following notation for functions of infinitely many variables. Let S_n be a metric space with metric $d_n(\cdot)$ and canonical point s_n or s_n'. On $S = \prod_n S_n$, with canonical point $s = (s_1, \cdots)$ or $s' = (s_1', \cdots)$, we use the metric $d(s, s') = \sum_n 2^{-n} d_n(s_n, s_n')/[1 + d_n(s_n, s_n')]$. Suppose that the number of control actions on each bounded time interval is finite. Then, from (3.12), we can construct a unique function $F(\cdot)$ with values in $D^2[0,\infty)$ and such that

$$(Y^{1,\varepsilon}(\cdot), Y^{2,\varepsilon}(\cdot)) = F(X_0^\varepsilon, W^\varepsilon(\cdot), B^\varepsilon(\cdot), X^{i,\varepsilon}(\mu_n^{i,\varepsilon}), \mu_n^{i,\varepsilon}, \tilde{\mu}_n^{i,\varepsilon}, \quad i = 1, 2,\ n = 1, 2, \cdots) \tag{3.13}$$

where $F(\cdot)$ is continuous (recall that we use the sup norm topology on bounded intervals on $D[0,\infty)$ here), and $Y^{i,\varepsilon}(\cdot)$ can increase only when $X^{i,\varepsilon}(t)$ equals zero. Also $X^{i,\varepsilon}(\cdot) \geqq 0$, always.

A tentative form for the limit control problem. To motivate the form of the limit process, suppose that the arguments of $F(\cdot)$ converge to $W(\cdot), B(\cdot), \cdots, \rho(\cdot)$, where $\rho(\cdot) = 0$, and let $Y^i(\cdot)$ be the limit of $Y^{i,\varepsilon}(\cdot)$. Then, on each bounded time interval, the complement of the intervals $\{[\mu_n^i, \tilde{\mu}_n^i), n < \infty]$ will just be a finite set of points, and the controls will be impulses acting at these points. Using this assumed convergence and the approximations in (3.9b), we can characterize the limit process as

$$\begin{aligned} X^1(t) &= X^1(0) + W^1(t) + B^1(t) + Y^1(t) - U^{01}(t) + U^1(t), \\ X^2(t) &= X^2(0) + W^2(t) + B^2(t) + Y^2(t) \\ &\quad - p_{12} Y^1(t)/(p_{12}+p_{10}) - U^{02}(t) - U_c^{12}(t). \end{aligned} \tag{3.14}$$

The sense will be made precise in Theorem 5.1.

The $Y_c^{12}(\cdot)$ can be obtained from the limit $Y^1(\cdot)$ via (3.9). The limits $Y^1(\cdot) = \lim_\varepsilon (Y^{10,\varepsilon}(\cdot) + Y^{12,\varepsilon}(\cdot))$ and $Y^2(\cdot) = \lim_\varepsilon Y^{20,\varepsilon}(\cdot)$ are obtained from the limit of (3.13). Furthermore (as in [1]), the $Y^i(\cdot)$ obtained from the limits in (3.13) are the unique continuous functions that can increase only when $X^i(t)$ is zero and that guarantee $X^i(t) \geqq 0$.

The $U_c^{12}(\cdot)$ can be used to define $U^{12}(\cdot)$ via the limits in (3.9). It will turn out that $U^{12}(\cdot) = p_{12}U^1(\cdot)/(1-p_{11})$.

4. Description of the limit control problem. In this section, we define the proper limit control problem for the system of Fig. 2.1. First, it will be convenient to describe the effects of various control actions on the $X^\varepsilon(\cdot)$ for small ε. We do this in some detail, since the limit problem is somewhat nonstandard, owing to the possibility of "multiple simultaneous impulses."

For the "limit" problem to make sense as an approximation to the physical problem, for any admissible policy π for $X(\cdot)$, there must be a sequence π^ε of policies that can be applied to the $X^\varepsilon(\cdot)$ (e.g., P_{ij}, P_i on/off) and such that, under π^ε, $X^\varepsilon(\cdot)$ converges to $X(\cdot)$ under policy π, and the associated costs also converge. Because of this, the limit control problem must be defined in terms of limits of what is possible for the $X^\varepsilon(\cdot)$.

Controls for the limit problem. Refer to Fig. 4.1, where some typical paths are constructed, under the heavy traffic conditions. Start at point (a) with all P_i, P_{ij} on except that P_{01} is off. The path moves to the left and as $\varepsilon \to 0$, it converges to the horizontal line (a, b). The mean (interpolated) movement to the left in time Δ is $g_{ai}\Delta/\sqrt{\varepsilon} + O(\Delta)$. Hence in the limit, as $\varepsilon \to 0$, there is an impulsive change.

Now, restart at (d) with only P_{12} off. The path drops, and as $\varepsilon \to 0$ it tends to the vertical line (d, e). In time Δ, the mean drop is $p_{12}g_{d1}\Delta/\sqrt{\varepsilon} + O(\Delta)$. The same path is followed if only P_{02} is off or if P_1 and P_{01} are both off, although the "drop" speed will be different. Now, restart at (e) with only P_1 off. The path moves toward (f) (for

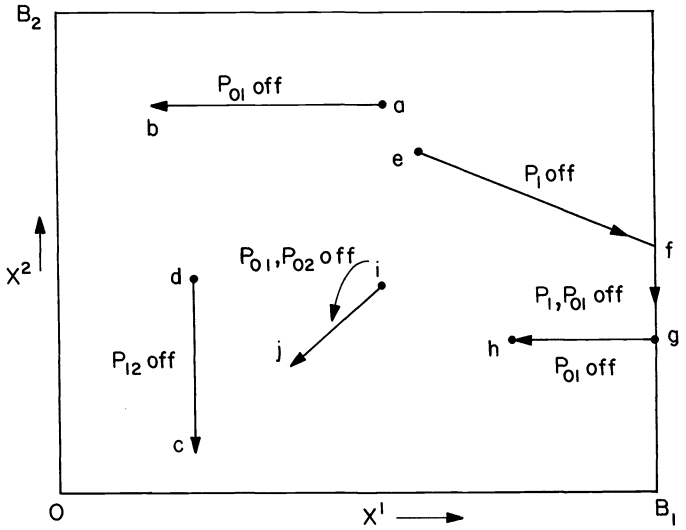

FIG. 4.1. *Impulsive changes in $X^\varepsilon(\cdot)$ or $X(\cdot)$ due to the control actions.*

small ε), and the limit slope can be calculated from

$$\frac{\text{net mean flow into } P_2}{\text{net mean flow into } P_1} = \frac{g_{a2}-(1-p_{22})g_{d2}}{g_{a1}} = -\frac{p_{12}g_{d1}}{g_{a1}}. \tag{4.1}$$

If the path reaches (f), then P_{01} must be turned off. If, at (g), we turn P_1 back on (but leave P_{01} off), then the path moves toward (h). The effects of both P_1 and P_{12} being off simultaneously are the same as for P_1 being off alone. Over small intervals of length Δ, the $\tilde{A}$, $\tilde{D}$, and Y terms in (3.10) contribute very little to the paths (compared to the effects of the control actions), since they converge weakly to continuous functions.

Now refer to (i), and let only P_{01} and P_{02} be off. Then the path moves to (j) with a limit slope $[(1-p_{22})g_{d2}-p_{12}g_{d1}]/(1-p_{11})g_{d1}$.

All finite sequences of arbitrary lengths of the impulses described in connection with Fig. 4.1 are possible. Suppose $(e)\to(f)\to(g)\to(h)$. Then as $\varepsilon\to 0$, it would appear that the limit $X(\cdot)$ jumps from (e) to (h) directly. But this $(e)\to(h)$ impulse must be realized as a concatenation of the basic impulses described above. In general the limit control is specified by a sequence of off/on actions for the P_i, P_{ij}, in a *specified order*, and with the impulsive distance traveled between successive ("simultaneous") control actions specified. The cost paid for the impulses is precisely the impulsive costs defined by (2.9). The described limitation on the ways in which the impulses for $X(\cdot)$ can be created is important, if the control problem for the limit $X(\cdot)$ is to be properly related to that for $X^\varepsilon(\cdot)$. In § 6, we show that the problem can be quite tractable from a numerical point of view.

The instantaneous changes in the $U^\alpha(\cdot)$ can be readily read off from the limit sequences of simultaneous impulses. For illustration, we do it for the (e, f, g, h) sequence of Fig. 4.1. Let e_i, etc. denote the ith coordinate of the point (e), and let δU^α denote the increment in U^α. On (e, f), $\delta U^{10}+\delta U^{12}=f_1-e_1$, $\delta U_c^{12}=e_2-f_2$. On (f, g), $\delta U^{01}=\delta U^{10}+\delta U^{12}$, and the value is unimportant, since their effects cancel in (2.8a). Also, $\delta U_c^{12}=f_2-g_2$. On (g, h), $\delta U^{01}=g_1-h_1$. All nonspecified δU^α are zero. The δU^{1i} always occur together as the sum $(\delta U^{10}+\delta U^{12})$.

The limit dynamical system. The Wiener process. The limit system will be (3.14), where the $W^i(\cdot)$ can be written in terms of the limits of the terms in (3.10) that are used to define them:

$$W^1(\cdot)=\tilde{A}^1(\cdot)+W_d^1(\cdot),\ W_d^1(\cdot)=-\tilde{D}^{10}(\cdot)-\tilde{D}^{12}(\cdot), \tag{4.2}$$

$$W^2(\cdot)=\tilde{A}^2(\cdot)+W_d^2(\cdot),\ W_d^2(\cdot)=-\tilde{D}^{20}(\cdot)+\tilde{D}^{12}(\cdot). \tag{4.3}$$

Here, all the terms are continuous martingales, with $\tilde{A}^1(\cdot)$, $\tilde{A}^2(\cdot)$, $\tilde{D}^{20}(\cdot)$ and $(\tilde{D}^{10}(\cdot), \tilde{D}^{12}(\cdot))$ being mutually orthogonal. The quadratic variation of $\tilde{A}^i(\cdot)$ is $\int_0^t g_{ai}^3\sigma_{ai}^2(X(s))\,ds$ and that of $W_d(\cdot)=(W_d^1(\cdot), W_d^2(\cdot))$ is $\Sigma(t)=\{\Sigma_{ij}(t)\}$, where

$$\begin{aligned}
\Sigma_{11}(t) &= g_{d1}\left[p_{11}(1-p_{11})t+g_{d1}^2(1-p_{11})^2\int_0^t \sigma_{d1}^2(X(s))\,ds\right],\\
\Sigma_{12}(t) &= -g_{d1}^3p_{12}(1-p_{11})\int_0^t \sigma_{d1}^2(X(s))\,ds-p_{12}p_{11}g_{d1}t,\\
\Sigma_{22}(t) &= g_{d2}\left[p_{20}(1-p_{20})t+p_{20}^2g_{d2}^2\int_0^t \sigma_{d2}^2(X(s))\,ds\right]\\
&\quad+g_{d1}\left[p_{12}(1-p_{12})t+p_{12}^2g_{d1}^2\int_0^t \sigma_{d1}^2(X(s))\,ds\right].
\end{aligned} \tag{4.4}$$

If the σ_{di}^2 and σ_{ai}^2 are constants, then the covariance is precisely that obtained by Reimann [1] (with a different notation used there).

It is evident from (4.4) and the cited orthogonality properties that there are mutually independent Wiener processes $w_a^i(\cdot)$, $w_d^i(\cdot)$, $w_d^{20}(\cdot)$, $\{w_d^{11}(\cdot), w_d^{12}(\cdot)\}$, where each scalar valued process is standard, and with respect to which $X(\cdot)$ is nonanticipative and $Ew_d^{11}(t)w_d^{12}(t) = -[p_{11}p_{12}/(1-p_{11})(1-p_{12})]^{1/2}t$ and

$$\tilde{A}^i(t) = g_{ai}^{3/2}\int_0^t \sigma_{ai}(X(s))\,dw_a^i(s),$$

$$W_d^1(t) = [g_{d1}p_{11}(1-p_{11})]^{1/2}w_d^{11}(t) + (1-p_{11})g_{d1}^{3/2}\int_0^t \sigma_{d1}(X(s))\,dw_d^1(s), \tag{4.5}$$

$$W_d^2(t) = [g_{d2}p_{20}(1-p_{20})]^{1/2}w_d^{20}(t) + [g_{d1}p_{12}(1-p_{12})]^{1/2}w_d^{12}(t)$$
$$+ p_{20}g_{d2}^{3/2}\int_0^t \sigma_{d2}(X(s))\,dw_d^2(s) - p_{12}g_{d1}^{3/2}\int_0^t \sigma_{d1}(X(s))\,dw_d^1(s).$$

The drift terms $B^i(\cdot)$ in (3.14) came from (3.6) and (3.7) and are

$$B^1(t) = \int_0^t [a^1(X(s)) - (1-p_{11})d^1(X(s))]\,ds,$$
$$B^2(t) = \int_0^t [a^2(X(s)) - (1-p_{22})d^2(X(s)) + p_{12}d^1(X(s))]\,ds. \tag{4.6}$$

Admissible control actions. The U_c^α and U^α in (3.14) are nondecreasing piecewise constant functions having only a finite number of jumps on each finite interval, and they can be taken to be right continuous. They thus correspond to "impulsive" controls. The allowed impulsive effects of U^1 in (3.14) are those described for $U^{1,\varepsilon}$ in (2.8), as $\varepsilon \to 0$. Also the impulsive effects of U_c^{12} are the limits of those of $U_c^{12,\varepsilon}$, and the effects of the U^{0i} are those of the $U^{0i,\varepsilon}$ as $\varepsilon \to 0$. This completely characterizes the possibilities for the impulse control of (3.14). Generally, several components of the controls might jump simultaneously, or a single jump in one component might be a consequence of a multiple simultaneous off/on sequence. We must allow these possibilities and distinguish an order for the "simultaneity," as discussed above, not only because they are possible control actions, but because they are possible limits of control actions for the physical processes. Thus, we count the parts of the multiple simultaneous impulses as distinct impulses. We now develop the notation for keeping track of the necessary information. Recall the definitions of τ_n^ε and R_n^ε given below (2.9).

Let τ_n denote the sequence of event times. The τ_n are not necessarily distinct, but $\tau_{n+1} \geqq \tau_n$ and the subscript n denotes the correct ordering, "simultaneous" or not. At each event time one or more of P_i or P_{ij} might shut off or turn on. What happens is indicated by the vector $R_n = (R_n^{01}, R_n^1, R_n^{02}, R_n^{12})$, where $R_n^{ij} = 1, -1$ or 0 (respectively, R_n^1) according to whether or not P_{ij} (respectively, P_1) is turned on, off, or not changed at τ_n. *Associated with* (τ_n, R_n) *is* $\delta U_n = (\delta U_n^{01}, \delta U_n^{02}, \delta U_n^1, \delta U_{c,n}^{12})$, *the instantaneous* (*at* τ_n) *change in the controls* $U(\cdot)$. To illustrate the procedure refer to the path (e, f, g, h) in Fig. 4.1. There are four event times: τ_1 associated with (e), τ_2 with (f), etc. Also $\tau_1 = \tau_2 = \tau_3 = \tau_4$. At τ_1, $R^1 = 1$. At τ_2, $R^{01} = 1$. At τ_3, $R^1 = -1$ and at τ_4, $R^{01} = -1$. All nonlisted R^α are zero. The associated impulses δU_n are given in the discussion below (4.3).

The $\{\delta U_n, \tau_n, R_n\}$ is said to be a control policy. The policy is said to be *admissible* if the function

$$\hat{\mathcal{R}}(t) = \{X_0, \delta U_n I_{\{\tau_n \leqq t\}}, \tau_n I_{\{\tau_n \leqq t\}}, R_n I_{\{\tau_n \leqq t\}}, I_{\{\tau_n \leqq t\}}, n < \infty, X(t), Y(t)\} \tag{4.7}$$

is nonanticipative with respect to the Wiener processes $w_\beta^\alpha(\cdot)$. An equivalent definition of admissibility is if the $\tilde{A}^i$, $\tilde{D}^{ij}(\cdot)$ are martingales with respect to the filtration generated by $\{\hat{\mathcal{R}}(t), \tilde{A}^i(\cdot), \tilde{D}^{ij}(\cdot)\}$, with the quadratic variation defined in and above (4.4).

The $Y(\cdot)$ in (3.14) is obtained from (5.1) below that is in turn obtained by taking limits in (3.13).

For an admissible policy, the cost function (the limit of (2.9)) is

$$
\begin{aligned}
V(\pi, x, P) = E_x^\pi \int_0^\infty e^{-\beta t} k(X(t))\, dt + k_1 E_x^\pi \sum_n e^{-\beta v_n^i} \\
+ \sum_1^2 k_{0i} E_x^\pi \sum_n e^{-\beta v_n^{0i}} + k_{12} E_x^\pi \sum_n e^{-\beta v_n^{12}} \\
+ E_x^\pi \int_0^\infty e^{-\beta t} \left[\sum_{i=1}^2 q_{0i}\, dU^{0i}(t) + q_{12}\, d[U_c^{12}(t) - U^{12}(t)] \right].
\end{aligned} \tag{4.8}
$$

In (4.8), the v_n^{ij}, v_n^i are defined as the moments of shutting off/turning on the indicated links or processors, as in § 2.

5. Weak convergence. Throughout the section, we use the Skorokhod topology on the products of $D[0, \infty)$, and the Euclidean topology on the Euclidean spaces.

We will use the following assumption:

(A5.1) The uncontrolled $X(\cdot)$ has a unique solution (in the weak sense) for each initial condition.

Note that (A5.1) implies weak uniqueness of the solution $X(\cdot)$ for any admissible control policy. Lemmas 5.1 and 5.2 are preparatory for the main convergence Theorem 5.1.

In the ensuing analysis, we ignore the possible increase of $Y_n^{1,\varepsilon}$ when $P_n^{12,\varepsilon} = 0$ (and $P_n^{1,\varepsilon} = 1$) for simplicity. This does not affect the result for the following reason. Shutting off the link from P_1 to P_2 does not affect the input or output process from P_1. Also, the mean number of times that the link can be shut off on any bounded time interval is bounded uniformly in ε, for otherwise the cost will go to infinity, as $\varepsilon \to 0$. Furthermore, the total mean interpolated time that the link is shut off on any bounded time interval goes to zero as $\varepsilon \to 0$, for otherwise the "$Z^{12,\varepsilon}(\cdot)$ component" of the cost will go to infinity as $\varepsilon \to 0$. A heavy traffic analysis of P_1 by itself yields a continuous limit $Y^1(\cdot)$ (see Lemma 5.3). The above facts imply that the mean increase of $Y^{1,\varepsilon}(\cdot)$ during all the intervals (on any $[0, T]$) when $P_n^{12,\varepsilon} = 0$ (and $P_n^{1,\varepsilon} = 1$) must go to zero as $\varepsilon \to 0$.

LEMMA 5.1. *Assume* (A2.1)-(A2.6) *and* (A5.1) *and let* $\sup_\varepsilon V^\varepsilon(\pi^\varepsilon, X_0^\varepsilon) < \infty$ *for* $\pi^\varepsilon = \{R_n^\varepsilon, \tau_n^\varepsilon, \delta u_n^\varepsilon, n < \infty\}$ *admissible. Then the first four convergences in* (3.9b) *all hold. On each interval* $[0, t]$ *the mean number of control actions is bounded uniformly in* ε *and each control interval collapses to a point as* $\varepsilon \to 0$.

Proof. We prove (3.9b) only for the process $\tilde{U}^{10,\varepsilon}(\cdot) = U^{10,\varepsilon}(\cdot) - p_{10} U^{1,\varepsilon}(\cdot)/(p_{10} + p_{12})$, since the rest are treated in the same way. Due to the mutual independence of the $\{I_n^{ij,\varepsilon}, n < \infty\}$ and its independence of $\{\psi_n^{1,\varepsilon}, P_n^1\}$,

$$
\tilde{U}^{10,\varepsilon}(t) = \sqrt{\varepsilon} \sum_1^{t/\varepsilon} \frac{[I_n^{10} p_{12} - I_n^{12} p_{10}]}{(p_{10} + p_{12})} \psi_n^1 (1 - P_n^1)
$$

is a martingale and its variance is bounded by $O(\varepsilon) E \sum_1^{t/\varepsilon} (1 - P_n^1) = C^\varepsilon(t)$. It is easily seen that $\overline{\lim}_\varepsilon \sqrt{\varepsilon}\, E \sum_1^{t/\varepsilon} (1 - P_n^1) < \infty$, for otherwise the buffer of P_1 will fill up (one or more times), forcing P_{01} to shut off (one or more times) such that $\overline{\lim}_\varepsilon E\, U^{01,\varepsilon}(t)$

and the associated costs will go to infinity. Thus $C^\varepsilon(t) \to 0$ as $\varepsilon \to 0$, yielding the desired result that $\tilde{U}^{10,\varepsilon}(\cdot)$ converges weakly to the zero process. The last assertion is obvious. □

Define $X_n^{i,\varepsilon}(\cdot) = X^{i,\varepsilon}((\mu_n^{i,\varepsilon} + \cdot) \cap \tilde{\mu}_n^{i,\varepsilon})$. Then (the $\delta_{in}^\varepsilon W$, etc., are defined above (3.12)) we can write

$$
\begin{aligned}
X_n^{1,\varepsilon}(t) &= X^{1,\varepsilon}(\mu_n^{1,\varepsilon}) + \delta_{1n}^\varepsilon W^{1,\varepsilon}(t) + \delta_{1n}^\varepsilon B^{1,\varepsilon}(t) + \delta_{1n}^\varepsilon \rho^{1,\varepsilon}(t) + \delta_{1n}^\varepsilon Y^{1,\varepsilon}(t), \\
X_n^{2,\varepsilon} &= X^{2,\varepsilon}(\mu_n^{2,\varepsilon}) + \delta_{2n}^\varepsilon W^{2,n}(t) + \delta_{2n}^\varepsilon B^{2,n}(t) + \delta_{2n}^\varepsilon \hat{\rho}^{2,\varepsilon}(t) \\
&\quad - p_{12}/(p_{12} + p_{10}) \delta_{2n}^\varepsilon Y^{1,\varepsilon}(t) + \delta_{2n}^\varepsilon Y^{2,\varepsilon}(t).
\end{aligned} \tag{5.1}
$$

In (5.1), $\{\mu_n^{1,\varepsilon}\}$ is the subset of $\{\tau_n^\varepsilon\}$ at which both P_1 and P_{01} are on, with at least one turned off at τ_{n-1}^ε, and $\{\mu_n^{2,\varepsilon}\}$ is the subset of times at which all of P_1, P_{12}, and P_{02} are on, with at least one being off at τ_{n-1}^ε.

LEMMA 5.2. *Assume the conditions of Lemma* 5.1. *For $\alpha = a$ and d, the processes $\varepsilon S_{\alpha, t/\varepsilon}^{i,\varepsilon}$ and $\bar{S}_\alpha^{i,\varepsilon}(\cdot)$ converge weakly to $S_\alpha^1(\cdot)$ and $\bar{S}_\alpha^i(\cdot)$, respectively, where $S_\alpha^i(t) = t/g_{\alpha i}$ and $\bar{S}_\alpha^i(t) = t g_{\alpha i}$. The $\{\tilde{A}^{1,\varepsilon}(\cdot), \tilde{A}^{2,\varepsilon}(\cdot), (\tilde{D}^{1,\varepsilon}(\cdot), \tilde{D}^{2,\varepsilon}(\cdot)), \tilde{D}^{20}(\cdot)\}$ is tight in $D^5[0,\infty)$ and the limits of any weakly convergent subsequence of the four sets (we pair $\tilde{D}^{10}$ and $\tilde{D}^{12}$) are strongly orthogonal continuous martingales. The $\{\delta_{in}^\varepsilon Y^{i,\varepsilon}(\cdot), \varepsilon > 0\}$ are tight and the weak limits are continuous, and similarly for the $\{Y^{i,\varepsilon}(\cdot), \varepsilon > 0\}$. The $\{X_n^{i,\varepsilon}(\cdot), \varepsilon > 0\}$ are tight and have continuous limits. The last two parts of* (3.9b) *hold, and the $\rho^{i,\varepsilon}(\cdot)$ and $\hat{\rho}^{i,\varepsilon}(\cdot)$ in* (3.10), (3.11), *all converge to the zero process.*

Proof. We now prove the first assertion and do only one case. We have

$$
\varepsilon S_{a,t/\varepsilon}^{i,\varepsilon} - S_a^i(t) = \varepsilon \sum_1^{t/\varepsilon} [\alpha_n^{i,\varepsilon} - \bar{\alpha}_n^{i,\varepsilon}] + O(\sqrt{\varepsilon}).
$$

The summands are martingale differences, since they are centered about their conditional expectations, given the "past." Thus, the variance of the sum is $O(\varepsilon)$. The weak convergence follows from this.

Owing to the above results, for the second assertion of the lemma we need only prove that the set $\{\tilde{A}_0^{1,\varepsilon}(\cdot), \tilde{A}_0^{2,\varepsilon}(\cdot), (\tilde{D}_0^{10,\varepsilon}(\cdot), \tilde{D}_0^{12,\varepsilon}(\cdot)), \tilde{D}_0^{20,\varepsilon}(\cdot)\}$ is tight and that the limits are continuous martingales, whose quadratic covariation = 0.

By the fact that the terms in the summands in each sum in the above set are centered about their conditional expectations (given the "past"), and the square integrability (A2.3), each term is a sum of martingale differences and is tight in the Skorokhod topology. This follows from the Aldous criterion [5, Thm. 3.3, Part 2].

By the uniform integrability in (A2.3), for each $T < \infty$, $\delta > 0$,

$$
P\left\{ \sup_{k \leq T/\varepsilon} |\sqrt{\varepsilon}(1 - \alpha_k^i / \bar{\alpha}_k^i)| > \delta \right\} \to 0.
$$

Then the jumps in $\tilde{A}_0^{i,\varepsilon}(\cdot)$ are "uniformly small." This implies that any weak limit of $\{\tilde{A}_0^{i,\varepsilon}(\cdot), \varepsilon > 0\}$ must have continuous paths w.p.1. This is similar for the other terms.

We next prove that all weak limits are orthogonal martingales. We first show the orthogonality of $\tilde{D}_0^{ij,\varepsilon}(\cdot)$ and $\tilde{A}_0^{i,\varepsilon}(\cdot)$. Take a "typical" term from each sum and use the definition of $E_{\alpha,n}^{i,\varepsilon}$ above (A2.2) and the centering in (3.1) to get (drop the ε for simplicity)

$$
\begin{aligned}
E\left[I_k^{ij} - p_{ij} \frac{\Delta_k^i}{\bar{\Delta}_k^i} \right] \left[1 - \frac{\alpha_n^i}{\bar{\alpha}_n^i} \right] &= E\left[I_k^{ij} - p_{ij} \frac{\Delta_k^i}{\bar{\Delta}_k^i} \right] I_{\{S_{d,k-1}^{i,\varepsilon} \leq S_{a,n-1}^{i,\varepsilon}\}} E_{a,n-1}^{i,\varepsilon} \left(1 - \frac{\alpha_n^i}{\bar{\alpha}_n^i} \right) \\
&\quad + E\left[1 - \frac{\alpha_n^i}{\bar{\alpha}_n^i} \right] I_{\{S_{a,n-1}^{i,\varepsilon} < S_{d,k-1}^{i,\varepsilon}\}} E_{d,k-1}^{i,\varepsilon} \left(I_k^{ij} - p_{ij} \frac{\Delta_k^i}{\bar{\Delta}_k^i} \right) = 0.
\end{aligned}
$$

By a similar calculation we can show the following. Let $h(\cdot)$ be a bounded and continuous function of its argument and let t_i, $i=1,\cdots,k$, t and s be such that $t_i \leqq t < t+s$. Then

$$Eh(\tilde{A}_0^{i,\varepsilon}(t_j), \tilde{D}_0^{ij,\varepsilon}(t_j), j \leqq k)[\tilde{A}_0^{i,\varepsilon}(t+s)\tilde{D}_0^{ij,\varepsilon}(t+s) - \tilde{A}_0^{i,\varepsilon}(t)\tilde{B}_0^{i,\varepsilon}(t)] = 0.$$

If $\tilde{A}_0^i(\cdot)$, $\tilde{D}_0^{ij}(\cdot)$ are weak limits, then

$$Eh(\tilde{A}_0^i(t_j), \tilde{D}_0^{ij}(t_j), j \leqq k)[\tilde{A}_0^i(t+s)\tilde{D}_0^{ij}(t+s) - \tilde{A}_0^i(t)\tilde{D}_0^{ij}(t)] = 0.$$

Due to the arbitrariness of $h(\cdot)$, k, t_j, t, $t+s$, this expression implies that $\tilde{A}_0^i(\cdot)$ and $\tilde{D}_0^{ij}(\cdot)$ are strongly orthogonal martingales. (The fact that they are martingales can be seen by repeating the argument and dropping one of the processes.) This argument yields the third sentence of the theorem, when applied to all the processes there.

The $\rho^{i,\varepsilon}(\cdot)$ satisfy $E\sup_{1\leqq T}|\rho^{i,\varepsilon}(t)| \overset{\varepsilon}{\to} 0$ for each T. The fact that $\{\delta_{1n}^{\varepsilon} Y^{i,\varepsilon}(\cdot), n=1,2,\cdots, \varepsilon>0\}$ is tight and has continuous limits follows from the above assertions concerning tightness and continuous limits of the processes in the arguments of the first part of (3.12). The assertion concerning $\{Y^{1,\varepsilon}(\cdot), \varepsilon>0\}$ follows from this argument and the fact that there are only finitely many control actions w.p.1 on each bounded time interval.

We now show that $(p_{10}+p_{12})Y^{12,\varepsilon}(\cdot) - p_{12}Y^{1,\varepsilon}(\cdot)$ converges to the zero process as $\varepsilon \to 0$. This expression equals $\sqrt{\varepsilon}\sum_1^{t/\varepsilon}[p_{10}I_n^{12} - p_{12}I_n^{10}]P_n^1 I_{\{X_n^1=0\}}$, which is a sum of martingale differences and has variance $O(\varepsilon)E\sum_1^{t/\varepsilon} I_{\{X_n^1=0\}}$. The fact that $Y^{1,\varepsilon}(\cdot)$ converges weakly to a continuous process implies that

$$P\left\{O(\varepsilon)\sum_1^{T/\varepsilon} I_{\{X_n^1=0\}} \geqq \delta\right\} \overset{\varepsilon}{\to} 0$$

for any $T>0$, $\delta>0$. The tightness and continuity of the weak limits of $\{\delta_{2n}^{\varepsilon} Y^{2,\varepsilon}(\cdot), n=1,2,\cdots, \varepsilon>0\}$ and of $\{Y^{2,\varepsilon}(\cdot), \varepsilon>0\}$ follows from the above assertions and the representation in the second line of (3.12). □

The following lemma is a corollary of Lemma 5.2.

LEMMA 5.3. *Assume the conditions of Lemma* 5.1. *Let ε index a weakly convergent subsequence of $\{X_n^{i,\varepsilon}(\cdot), \mu_n^{i,\varepsilon}, \tilde{\mu}_n^{i,\varepsilon}, W^{i,\varepsilon}(\cdot), \varepsilon>0, i=1,2, n=1,2,\cdots\}$ with limit denoted by $\{X_n^i(\cdot), \mu_n^i, \tilde{\mu}_n^i, W^i(\cdot), i=1,2, n=1,2,\cdots\}$. Then $\mu_{n+1}^i = \tilde{\mu}_n^i$ and $X_n^i(0) = \lim_{\varepsilon} X^{i,\varepsilon}(\mu_n^{i,\varepsilon})$. For a real-valued function $G(\cdot)$ on $[0,\infty)$, define the function $\delta_{in}G(\cdot) = G((\mu_n^i+\cdot)\cap\mu_{n+1}^i) - G(\mu_n^i)$. Then $\{Y^{i,\varepsilon}(\cdot), i=1,2, \varepsilon>0\}$ converges weakly to $(Y^1(\cdot), Y^2(\cdot))$, where $Y^i(\cdot) = \sum_{n:\mu_n^i\leqq t}\delta_{in}Y^i(t-\mu_n^i)$ and $\delta_{in}Y^i(\cdot)$ is just the weak limit of $\{\delta_{in}^{\varepsilon}Y^{i,\varepsilon}(\cdot), \varepsilon>0\}$. We have*

$$X_n^1(t) = X_n^1(0) + \delta_{1n}W^1(t) + \delta_{1n}B^1(t) + \delta_{1n}Y^1(t),$$

$$X_n^2(t) = X_n^2(0) + \delta_{2n}W^2(t) + \delta_{2n}B^2(t) + \delta_{2n}Y^2(t) - p_{12}\delta_{2n}Y^1(t)/(p_{12}+p_{10}).$$

The $\delta_{in}Y^i(\cdot)$ are continuous and can increase only at those times when $X_n^i(\cdot)$ equals zero. Also $X_n^i(t) \geqq 0$, for all t, n, i.

The proof follows from Lemma 5.2, the representation (3.12), and the weak convergence.

Notation for Theorem 5.1. Let $\mathscr{R}^{\varepsilon}$ denote the set $\{X_0^{\varepsilon}, \tilde{A}^{i,\varepsilon}(\cdot), \tilde{D}^{ij,\varepsilon}(\cdot), B^{\varepsilon}(\cdot), R_n^{\varepsilon}(\cdot), \tau_n^{\varepsilon}, \delta U_n^{\varepsilon}, i,j,n\}$, where the $\mathscr{R}$, τ and δU are defined below (2.9). If ε indexes a weakly convergent subsequence of $\{\mathscr{R}^{\varepsilon}\}$ with limit $\mathscr{R}$, we use the following notation.

Define the process

$$\mathscr{R}(\cdot) = \{X_0, \tilde{A}^i(\cdot), \tilde{D}^{ij}(\cdot), B(\cdot), i, j, (R_n, \tau_n, \delta U_n)I_{\{\tau_n\leqq\cdot\}}, n<\infty\},$$

and let $\mathcal{B}(t)$ denote the σ-algebra induced by $\{\mathcal{R}(s), s \leqq t\}$. Analogously to the definition of R_n^ε and δU_n^ε, we write $R_n = (R_n^1, R_n^{01}, R_n^{02}, R_n^{12})$ and $\delta U_n = (\delta U_n^1, \cdots, \delta U_{cn}^{12})$. Define $U^\alpha(t) = \sum_{n:\tau_n \leqq t} \delta U_n^\alpha$. The $\{\mu_n^i\}$ is a subset of $\{\tau_n\}$.

THEOREM 5.1. *Assume the conditions of Lemma* 5.1. *Then* $\{\mathcal{R}^\varepsilon\}$ *is tight. Let* ε *index a weakly convergent subsequence with limit* $\mathcal{R}$. *Let* $X(\cdot)$ *denote the process with paths in* $D[0, \infty)$ *that equals* $X_n^i(t - \mu_n^i)$ *for* $t \in [\mu_n^i, \mu_{n+1}^i)$. *Then* $(X^1(\cdot), X^2(\cdot))$ *satisfy* (3.14), *where* $W^1(\cdot) = \tilde{A}^1(\cdot) - (\tilde{D}^{10}(\cdot) + \tilde{D}^{12}(\cdot))$ *and* $W^2(\cdot) = \tilde{A}^2(\cdot) - \tilde{D}^{20}(\cdot) + \tilde{D}^{12}(\cdot)$. *The* $Y^i(\cdot)$ *can increase only when* $X^i(\cdot)$ *takes the value zero, and* $X^i(t) \in [0, B_i]$. *Then* $\tilde{A}^i(\cdot)$ *and* $\tilde{D}^{ij}(\cdot)$ *are continuous* $\mathcal{B}(t)$*-martingales with quadratic variations given in and above* (4.4). *The limit policy* $\pi = \{R_n, \tau_n, \delta U_n\}$ *is admissible for* $X(\cdot)$.

Remark. We might not have $X^\varepsilon(\cdot) \Rightarrow X(\cdot)$ in the Skorokhod topology. The problem concerns the behavior during the "control intervals." For example, let $t_0 = \varepsilon k_\varepsilon > 0$ and consider the sequence of right continuous and piecewise constant functions defined by $M^\varepsilon(t) = 0$ for $t \leqq t_0$, and that then increases at each $t = \varepsilon k$ $(k \geqq k_0)$ by $\sqrt{\varepsilon}$ until the value of one is reached. In an obvious sense, the limit is a step function—with an increase of unity at t_0, but the convergence is not in the Skorokhod topology. We still get what is desired for our control problem.

Proof. By Lemmas 5.1 and 5.2, the processes and random variables in R^ε are tight, so that we can extract and work with a weakly convergent subsequence. Also, the limits of the "~" processes are continuous martingales. The assertion concerning $Y^i(\cdot)$ and the representation (3.14) follows from Lemma 5.3.

Owing to the strong orthogonality of the four processes $\tilde{A}^{i,\varepsilon}(\cdot)$, etc., it is sufficient to prove the "quadratic variation" property separately for each component. We do it only for $(\tilde{D}^{10}(\cdot), \tilde{D}^{12}(\cdot))$. Let ε index a weakly convergent subsequence of $\{\mathcal{R}^\varepsilon\}$. Let $f(\cdot)$ be a smooth function with compact support and $h(\cdot)$ a bounded and continuous function, both being real valued. Let the t, $t+s$ and $t_k \leqq t$ below be points such that the probability $P\{\tau_n$ equals t or $t+s$ or $t_k\} = 0$ for each n, k. Define $\delta\psi_n^{ij,\varepsilon} = I_n^{ij,\varepsilon} - p_{ij}\Delta_n^{i,\varepsilon}/\bar{\Delta}_n^{i,\varepsilon}$.

By the uniform integrability (A2.3), the representation of $\tilde{D}^{ij,\varepsilon}(\cdot)$ as a sum of the $\delta\psi_n^{ij,\varepsilon}$, and a truncated Taylor series expansion, for each $N < \infty$ we can write

$$
\begin{aligned}
&Eh(X^\varepsilon(t_k), \tilde{A}^{i,\varepsilon}(t_k), \tilde{D}^{ij,\varepsilon}(t_k), B^{i,\varepsilon}(t_k), (R_n^\varepsilon, \tau_n^\varepsilon, \delta\bar{U}_n^\varepsilon) I_{\{\tau_n^\varepsilon \leqq t_k\}}, k, n \leqq N) \\
&\quad \cdot \Bigg[f(\tilde{D}^{10,\varepsilon}(t+s), \tilde{D}^{12,\varepsilon}(t+s)) - f(\tilde{D}^{10,\varepsilon}(t), \tilde{D}^{12,\varepsilon}(t)) \\
&\qquad - \sqrt{\varepsilon} \sum_{\alpha=0,2} \sum_{\varepsilon n = \bar{S}_d^{1,\varepsilon}(t)}^{\bar{S}_d^{1,\varepsilon}(t+s)} f_{x_\alpha}\left(\sqrt{\varepsilon} \sum_1^{n-1} \delta\psi_k^{10}, \sqrt{\varepsilon} \sum_1^{n-1} \delta\psi_k^{12}\right) \cdot \delta\psi_n^{1\alpha} \\
&\qquad - \frac{1}{2}\varepsilon \sum_{\alpha,\beta=0,2} \sum_{\varepsilon n = \bar{S}_d^{1,\varepsilon}}^{\bar{S}_d^{1,\varepsilon}(t+s)} f_{x_\alpha x_\beta}\left(\sqrt{\varepsilon} \sum_1^{n-1} \delta\psi_k^{10}, \sqrt{\varepsilon} \sum_1^{n-1} \delta\psi_k^{12}\right) \cdot \delta\psi_n^{1\alpha} \delta\psi_n^{1\beta} \Bigg] \\
&\quad \overset{\varepsilon}{\to} 0.
\end{aligned} \tag{5.2}
$$

Equation (5.2) holds since the conditional expectation of the bracketed quantity, given the data in the argument of $h(\cdot)$, goes to zero in the mean as $\varepsilon \to 0$.

By the properties of conditional expectations, (5.2) remains true if we replace each term of the sums in the bracket by its conditional expectation given any data that includes the data in the argument of $h(\cdot)$. Now, to exploit this, use the definition of $E_{d,n}^{1,\varepsilon}$ given above (A2.2), the centering of $\delta\psi_k^{ij,\varepsilon}$ and the assumption (A2.6) on the conditional variances, to replace $\delta\psi_k^{1\alpha}$ in the first sum in (5.2) by zero $(E_{d,k-1}^{1,\varepsilon}\delta\psi_k^{1\alpha} = 0)$,

and the $\delta\psi_k^{1\alpha}\delta\psi_k^{1\beta}$ in the second by $E_{d,k-1}^{1,\varepsilon}\delta\psi_k^{1\alpha}\delta\psi_k^{1\beta}$. For k any random time $\geqq \bar{S}_d^{1,\varepsilon}(t)$, this latter quantity is

$$E_{d,k-1}^{1,\varepsilon}\left[I_k^{1\alpha}-p_{1\alpha}\frac{\Delta_k^1}{\bar{\Delta}_k^1}\right]\left[I_k^{1\beta}-p_{1\beta}\frac{\Delta_k^1}{\bar{\Delta}_k^1}\right]=p_{1\alpha}\delta_{\alpha\beta}-p_{1\alpha}p_{1\beta}+p_{1\alpha}p_{1\beta}\,\mathrm{var}_{d,k-1}^{1,\varepsilon}\,\Delta_k^1/(\bar{\Delta}_k^1)^2$$

$$=p_{1\alpha}\delta_{\alpha\beta}-p_{1\alpha}p_{1\beta}+p_{1\alpha}p_{1\beta}g_{d1}^2\sigma_{d1}^2(X^{\varepsilon}_{S_{d,k-1}^{1,\varepsilon}})+(\text{negligible terms}). \tag{5.3}$$

With these replacements, the limit (as $\varepsilon\to 0$) of the double sum in (5.2) is

$$\frac{1}{2}\sum_{\alpha,\beta=0,2}\int_{\bar{S}_d^1(t)}^{\bar{S}_d^1(t+s)} f_{x_\alpha x_\beta}(\tilde{D}_0^{10}(\tau),\tilde{D}_0^{12}(\tau))\cdot\hat{\Sigma}_{\alpha\beta}(\tau)\,d\tau$$

where

$$\hat{\Sigma}_{00}(t)=p_{10}-p_{10}^2+p_{10}^2g_{d1}^2\sigma_{d1}^2(X(t/g_{d1})),$$
$$\hat{\Sigma}_{02}(t)=-p_{10}p_{12}+p_{10}p_{12}g_{d1}^2\sigma_{d1}^2(X(t/g_{d1})),$$
$$\hat{\Sigma}_{22}(t)=p_{20}-p_{20}^2+p_{20}^2g_{d1}^2\sigma_{d1}^2(X(t/g_{d1}))$$

where we used (5.3) and the fact that $\varepsilon S_{d,t/\varepsilon}^{1,\varepsilon}\to t/g_{d1}$ to get the proper limit of the argument of $\sigma_{d1}^2(\cdot)$. The right-hand sides are defined for all t that are not points of control action (i.e., for all but a finite number of t, w.p.1).

Now, recalling that $\bar{S}_d^i(t)=g_{di}t$, and taking limits in (5.2), we obtain

$$Eh(X(t_k),\tilde{A}^i(t_k),\tilde{D}^{ij}(t_k),B^i(t_k),(R_n,\tau_n,\delta U_n)I_{\{\tau_n\leqq t_k\}},n\leqq N,k\})$$
$$\cdot\Big[f(\tilde{D}^{10}(t+s),\tilde{D}^{12}(t+s))-f(\tilde{D}^{10}(\tau),\tilde{D}^{12}(t))$$
$$-\frac{1}{2}\sum_{\alpha,\beta=0,2}\int_{tg_{di}}^{(t+s)g_{di}} f_{x_\alpha x_\beta}(\tilde{D}^{10}(\tau),\tilde{D}^{12}(\tau))\hat{\Sigma}_{\alpha\beta}(\tau)\,d\tau\Big]=0. \tag{5.4}$$

The arbitrariness of $h(\cdot)$, $f(\cdot)$, N, t, $t+s$, and $\{t_i\}$ implies that the expectation of the bracketed term, conditional on $\mathcal{B}(t)$, is zero. Thus, (5.4) implies the asserted martingale property of $\tilde{D}^{1i}(\cdot)$.

The quadratic variation can be obtained from (5.4) by observing that $\tilde{D}_0^{1i}(t)=\tilde{D}^{1i}(g_{di}t)$ and using the change of variables $\tau/g_{d1}\to\tau$ and setting $f(x,y)$ to either x^2, xy, or y^2.

With analogous calculations for $\tilde{D}^{20,\varepsilon}(\cdot)$ and for the $\tilde{A}^{i,\varepsilon}(\cdot)$, we get the quadratic variation for the $W_\alpha^i(\cdot)$, $\tilde{A}^i(\cdot)$, $\tilde{D}^{ij}(\cdot)$, as given in § 4.

By the above argument, the limit policy $\{\tau_n,R_n,\delta U_n\}$ is "nonanticipative" with respect to the martingales. Owing to the way it was obtained as a limit of the $\{\tau_n^\varepsilon,R_n^\varepsilon,\delta U_n^\varepsilon\}$, the limit policy $\{\tau_n,R_n,\delta U_n\}$ is admissible in the sense that it corresponds to admissible sequences of impulses corresponding to the sequence of off/on controls as discussed in § 4. □

Extension. Consider the graph of $X^\varepsilon(\cdot)$ ($X^{1,\varepsilon}(\cdot)$ plotted versus $X^{2,\varepsilon}(\cdot)$) in the state space during a fixed control action. It can be shown that the graph converges uniformly (in probability) to the limit straight lines given by, for example, Fig. 4.1. The convergence is in the sense that the maximum value of the distance between any point on (this part of) the graph of $X^\varepsilon(\cdot)$ and the closest point on the limit straight line goes to zero in probability.

Alternatively, during the "control sections," use a rescaling with ε used for both the amplitude and timescale. Then the processes during this control interval converge in the Skorokhod topology, to straight lines whose graphs are precisely the graph referred to above.

THEOREM 5.2. *Assume* (A2.1)-(A2.6) *and* (A5.1), *and let* ε *index a weakly convergent subsequence with limit* $\mathcal{R}(\cdot)$. *Then* (*with* π *defined as in Theorem* 5.1) *for any* P

$$\underline{\lim_{\varepsilon}}\, V^{\varepsilon}(\pi^{\varepsilon}, x, P) \geqq V(\pi, x, P). \tag{5.5}$$

Define $N^{\alpha,\varepsilon}(t)$ *to be the number of actions of the control* P_α *on the interval* $[0, t]$. *If*

$$\{N^{\alpha,\varepsilon}(n+1) - N^{\alpha,\varepsilon}(n), \alpha, n < \infty\} \tag{5.6}$$

is uniformly integrable, then

$$V^{\varepsilon}(\pi^{\varepsilon}, x, P) \to V(\pi, x, P). \tag{5.7}$$

Proof. The relation (5.5) is just a consequence of Fatou's Lemma and the weak convergence. Now, let the uniform integrability hold. Then the holding costs and the impulsive control costs in (2.9) converge to their limits, as given by the terms in (4.8). We need only work with the last integral in (2.9). The arguments for each component are essentially the same, and we work with the $U^{01,\varepsilon}(\cdot)$ term only and assume that P_1 is on. If P_1 might also be off part of the time, the argument is a little more involved (involving the $X^{2,\varepsilon}$ as well as the $X^{1,\varepsilon}$), but is essentially the same.

When P_{01} is off, the increments in the $Y^{1,\varepsilon}(\cdot)$ are zero. (If $X^{1,\varepsilon}(t) = 0$, we must have P_{01} on, by (A2.1).) We can write

$$\begin{aligned} U^{01,\varepsilon}(t) &= \sum_n [U^{01,\varepsilon}(\tilde{v}_n^{01,\varepsilon} \cap t) - U^{01,\varepsilon}(v_n^{01,\varepsilon} \cap t)] \\ &= \sum_n [W^{1,\varepsilon}(\tilde{v}_n^{01,\varepsilon} \cap t) - W^{1,\varepsilon}(v_n^{01,\varepsilon} \cap t)] \\ &\quad - \sum_n [X^{1,\varepsilon}(\tilde{v}_n^{01,\varepsilon} \cap t) - X^{1,\varepsilon}(v_n^{01,\varepsilon} \cap t)] \\ &\quad + \sum_n [B^{1,\varepsilon}(\tilde{v}_n^{01,\varepsilon} \cap t) - B^{1,\varepsilon}(v_n^{01,\varepsilon} \cap t)] \\ &\quad + (\text{terms that go to zero as } \varepsilon \to 0). \end{aligned}$$

For some $K_1 < \infty$, the last two sums on the right are bounded by $K_1 N^{01,\varepsilon}(t)$, which is uniformly integrable by hypothesis. By the orthogonality properties of the summands in the expression for the $W^{1,\varepsilon}(\cdot)$, the mean square value of the middle term is $O(t+1)$. This yields the uniform integrability of $\{U^{01,\varepsilon}(t)\}$ for each t and of $\{U^{01,\varepsilon}(n+1) - U^{01,\varepsilon}(n), \varepsilon > 0, n < \infty\}$. By the weak convergence and the uniform integrability of these and the other terms in the last integral of (2.9), the assertion (5.7) follows. □

It is not a priori obvious that there is a control policy for which (5.6) is uniformly integrable, since we must shut off the inputs to P_i whenever its buffer is full. We will define a standard "comparison" control policy called the Δ_0-*boundary policy*, for which (5.6) is uniformly integrable. Since we can switch to this policy at any time, for any $\delta > 0$ there is a δ-optimal policy for which (5.6) is uniformly integrable. Let $\Delta_0 \in (0, \min(B_1, B_2)/4)$ and refer to Fig. 5.1. If $X^{2,\varepsilon} = B_2$ then shut off all inputs to P_2 until $X^{2,\varepsilon}$ reaches $B_2 - \Delta_0$. Then turn them back on. If at the end of that time $B_1 - \Delta_0 < X^{1,\varepsilon} \leqq B_1$, shut off P_{01} until $X^{1,\varepsilon} = B_1 - \Delta_0$. If $X^{1,\varepsilon} = B_1$, then shut off P_{01} until $X^{1,\varepsilon}$ reaches $B_1 - \Delta_0$. Then turn P_{01} back on. We use the analogous definition for the Δ_0-boundary policy for $X(\cdot)$. Then, if ever $X^{\varepsilon}(\cdot)$ or $X(\cdot)$ hits the outer boundary, we control it to a distance at least Δ_0 (in each coordinate) from the outer boundary.

THEOREM 5.3. *Assume* (A2.2)-(A2.6). *Then for the* Δ_0-*boundary control and each* $k < \infty$

$$\sup_{\substack{\varepsilon \text{ small} \\ \alpha, x, n}} E_x |N^{\alpha,\varepsilon}(n+1) - N^{\alpha,\varepsilon}(n)|^k < \infty \quad \textit{for all } \alpha, \tag{5.8}$$

and similarly for the "jump numbers" of the limit process $X(\cdot)$.

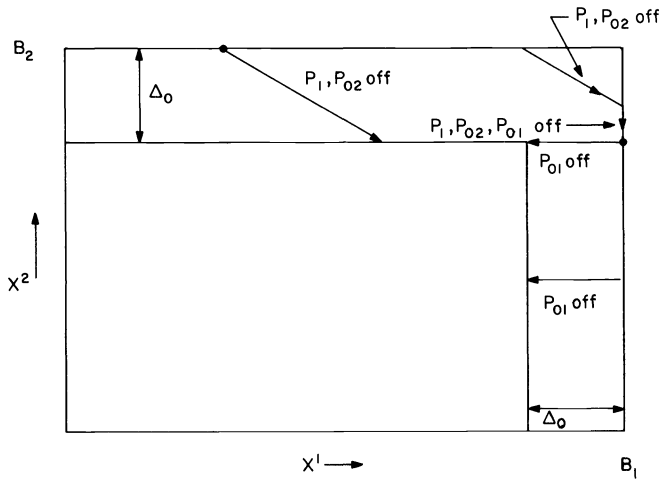

FIG. 5.1. *The comparison Δ_0-boundary control.*

Remark on the proof. Refer to Fig. 5.1. Let t_i^ε denote the ith time of return of $X^\varepsilon(\cdot)$ to the outer boundary after the ith time that the control takes the process to the set $[0, (B_1-\Delta_0)]\times[0, (B_2-\Delta_0)]$. We can readily show that for any $\delta_0 \in (0, 1)$, there is $T_0 > 0$ such that

$$\sup_{\substack{\omega,i,\\ \text{small } \varepsilon}} P\{t_{i+1}^\varepsilon - t_i^\varepsilon < T_0 \,|\, \text{data up to } t_i^\varepsilon\} \leqq 1-\delta_0. \tag{5.9}$$

This is just a consequence of the properties of $W^\varepsilon(\cdot)$ and $B^\varepsilon(\cdot)$ and of the fact that $dU^{\alpha,\varepsilon}(\cdot) = 0$ on the intervals of interest. With (5.9), it is not hard to show that all the moments of $N^{\alpha,\varepsilon}(iT_0+T_0) - N^{\alpha,\varepsilon}(iT_0)$ are bounded, uniformly in i and ε and in the initial condition (similarly, for the $X(\cdot)$ process). This yields the desired result. See the proof of Theorem 5.3 of [7] of a related result for a problem with a more complicated statistical structure.

The optimality and "almost" optimality theorem. At the present time almost nothing is known about optimal or δ-optimal ($\delta > 0$) policies for the $X^\varepsilon(\cdot)$. This is one of the basic reasons for considering suitably adapted policies that are "good" for $X(\cdot)$. Unfortunately, we know little about the optimal or δ-optimal policies for $X(\cdot)$. Thus, we must postulate (in (A5.2)) the existence of a δ-optimal policy with certain smoothness properties. The assumption appears to be eminently reasonable, since there is usually enormous flexibility in the smoothing that can be put on δ-optimal controls. The numerical results obtained via the methods described in § 6 satisfy (A5.2) for all the cases tried, in the sense that the "control decision" surfaces (discretized for the numerical calculation) seem to have the required properties. In fact, the situation in Fig. 5.1 seems to be typical, in the sense that some continuous deformation of these decision surfaces is what is seen in the numerical calculations.

For our current purposes, it is best to view the path $X(\cdot)$ as its graph in the state space. The uncontrolled sections are the graphs of the paths of the uncontrolled reflected diffusion, and the controlled sections are straight lines or "broken" straight lines, each segment corresponding to a different value of the set of indicators $P = (P^{01}, P^{02}, P^1, P^{12})$. In a sense, (A5.2) is a long-winded and formal way of saying that (for some δ-optimal policy) the lengths of the line segments are piecewise continuous in their starting point. It also deals with the possibility that the initial P might be

inappropriate for the initial state x, and that we might have to change the control settings instantaneously at $t=0$. We tried to give a general description of what reasonably seems to be expected. The situation might be simpler in special cases—but it seems likely that the useful δ-optimal control policies would be described by (A5.2), due to the nature of the impulse sequences. Note that

$$1+\sup_{x,P}[V(x,P)+1]/\min_{\alpha} k_\alpha \equiv \bar{K}$$

is an upper bound for the number of "simultaneous impulses" for the δ-optimal controls, with $\delta \leqq 1$.

We require some "smoothness" in the δ-optimal "feedback" controls, since we need to adapt them for use with the $X^\varepsilon(\cdot)$ process and will require that the corresponding sequence $\{X^\varepsilon(\cdot)\}$ (and the associated costs) converge appropriately to $X(\cdot)$ (and its associated cost).

The boundaries of the sets $G(1)$ and $G_i(P)$ below are smooth in that they are composed of a finite number of differentiable curves that are not tangent at the points of intersection. We use P to denote the control value just before a decision to change the control is made, and P_1 to denote the new control value just after the decision is made. Recall that $P=1$ is used for $P=(1,1,1,1)$.

We could replace (A5.2) by the simpler assumption that for each $\delta>0$ there is a δ-optimal admissible policy π_δ for $X(\cdot)$ and admissible policies π_δ^ε for $X^\varepsilon(\cdot)$ such that $X^\varepsilon(\cdot)$ (under π_δ^ε) converges weakly to $X(\cdot)$ (under π_δ), and the associated costs converge. Assumption (A5.2) simply defines a reasonable π_δ for which this can be done. The interiors of all sets in (A5.2) are relative to $G=[0,B_1]\times[0,B_2]$.

(A5.2) For each $\delta>0$, there is a policy π_δ for $X(\cdot)$ that is δ-optimal in the sense that it satisfies (A2.1) and

$$V(x,P)=\inf_{\pi \text{ adm.}} V(\pi,x,P) \geqq V(\pi_\delta,x,P)-\delta \tag{5.10}$$

for all x, P and that has the following properties:

(a) Let $P=1$. Then there is a decision set $G(1)$, whose boundary is divided into a finite number of segments. Each segment is associated with a switch to some $P_1 \neq 1$ when $X(\cdot)$ hits it from the exterior of $G(1)$. The segment associated with each P_1 is strictly interior to one of the sets $G_i(P_1)$ below.

(b) For each $P \neq 1$, there are a finite number (perhaps zero—see remark in (c) below) of sets $G_i(P)$ whose interiors are disjoint. If $x \in G_i(P)$ and P is used, then it is used until the boundary of $G_i(P)$ is reached. The distance (taken by the graph of $X(\cdot)$, a straight line) from $x \in G_i(P)$ to the exit point on the boundary of $G_i(P)$ is a continuous function of x. The (straight line) graph is (uniformly) not tangent to the boundary at any point of contact. The boundary is divided into a finite number of segments, each associated with a new control setting, perhaps with $P=1$. These segments are strictly interior to some set $G_j(P_1)$ for the new value P_1.

At the corners of the segments of $\partial G_i(P)$ or $\partial G(1)$, any policy associated with the intersecting segments can be used. There is $\Delta_1>0$ such that after a finite number of switches, we have $P=1$ and $X(\cdot)$ is a distance $\geqq \Delta_1$ from $G(1)$.

(c) It is possible that there will be immediate (or several) changes $P \to$ some $P_1 \neq P$ at $t=0$, until (a), (b) above are active.

Remark. The assumption concerning "points in common" to several $\partial G_i(P)$ does not seem to be restrictive. Generally, in dynamic programming, when the state is on the boundary of sets corresponding to different policies, any one of the policies is optimal.

Adapting π_δ to $X^\varepsilon(\cdot)$. By adapting the policy π_δ for use with $X^\varepsilon(\cdot)$ we simply take as the moments of decision the moments when $X^\varepsilon(\cdot)$ hits the decision boundary segments.

We now prove the "almost δ-optimality" of π_δ-applied to $X^\varepsilon(\cdot)$. This justifies the use of the limit approximations for purposes of getting nearly optimal controls.

THEOREM 5.4. *Assume* (A2.2)-(A2.6), (A5.1), *and* (A5.2). *Let π_δ^ε denote the policy of* (A5.2) *adapted to $X^\varepsilon(\cdot)$. Then*

$$V^\varepsilon(\pi_\delta^\varepsilon, x, P) \to V(\pi_\delta, x, P). \tag{5.11}$$

For admissible π^ε and small ε,

$$\sup_{\{\pi^\varepsilon\}} [V^\varepsilon(\pi_\delta^\varepsilon, x, P) - V^\varepsilon(\pi^\varepsilon, x, P)] \leqq 2\delta. \tag{5.12}$$

Proof. The proof is a consequence of the weak convergence in Theorems 5.1 and 5.3, the piecewise continuity properties of (A5.2) and an estimate of the type obtained in Theorem 5.2, and we only outline some of the argument.

(a) Let $\mathscr{R}^\varepsilon$ denote a sequence (as used in Theorem 5.1) associated with admissible π^ε. Let $\mathscr{R}_\delta^\varepsilon$ (respectively, $\mathscr{R}_\delta$) denote a convergent subsequence associated with π_δ^ε (its limit, respectively). Let $\bar{\mathscr{R}}_\delta$ and $\bar{X}_\delta(\cdot)$ denote the quantities associated with policy π_δ—with the understanding of the "multiple choices" at the corners in (A5.2)(b). Suppose that $N^\varepsilon(t)$ denotes the number of distinct control actions on $[0, t]$, and suppose that $\{N^{\alpha,\varepsilon}(n+1) - N^{\alpha,\varepsilon}(n),\ n < \infty,\ (\text{small})\ \varepsilon > 0\}$ is uniformly integrable for π_δ^ε. Now, suppose that if $\mathscr{R}_\delta$ is the limit of a subsequence of $\{\mathscr{R}_\delta^\varepsilon\}$, then $\bar{\mathscr{R}}_\delta = \mathscr{R}_\delta$. Then (5.11) holds.

The expression (5.12) is a consequence of the δ-optimality of π_δ and Theorem 5.2. Thus, we need only prove the uniform integrability and the fact that $\bar{\mathscr{R}}_\delta = \mathscr{R}_\delta$.

(b) The uniform integrability property is proved by a method similar to that used for the Δ_0-boundary policy.

(c) We examine only the case where $X(0) = X^\varepsilon(0) = x \notin G(1)$ and where there is no switch at $t = 0$, for the other cases are dealt with in a similar way. To get the desired weak convergence, we need to show that the hitting times and locations on the decision sets converge (as $\varepsilon \to 0$) to those that would hold under π_δ. Until the last remark below, we assume for simplicity that the hitting locations of the limits $X_\delta(\cdot)$ are not on the "corners." We first make the following observation.

Let H denote a compact set which is the closure of its interior and with differentiable boundary. Then for any $T < \infty$, the functions $\tau(x(\cdot)) = \min\{T, \text{hitting time of } x(\cdot) \text{ on } H\}$, $x(\tau(x(\cdot)))$ on $C[0, T]$ (sup norm) are continuous w.p.1 with respect to Wiener measure (similarly, if H satisfies the conditions that we put on $G(1)$). Let ε index a weakly convergent subsequence of $\{X^\varepsilon(\cdot)\}$ with limit denoted by $X_\delta(\cdot)$. We assume for simplicity that the hitting locations on the switching curves of $X_\delta(\cdot)$ are not on the corners (w.p.1). This is unrestrictive since by (A5.2) any of the actions that can be chosen at the corners yield the same cost. Then by the above "continuity" comment and the nondegeneracy of $(W^1(\cdot), W^2(\cdot))$, the first hitting times τ_1^ε of $X^\varepsilon(\cdot)$ on $G(1)$ converge weakly to τ_1, the first hitting time of $X_\delta(\cdot)$ on $G(1)$ and $X^\varepsilon(\tau_1^\varepsilon) \Rightarrow X_\delta(\tau_1)$.

Suppose that the first hitting point of $X_\delta^\varepsilon(\cdot)$ on $G(1)$ involves a switch to $P_1 \neq 1$, and is in the set $G_1(P_1)$. Let τ_2^ε denote the hitting time of the next decision set. Then (see comments after the proof of Theorem 5.1) the graph of $X^\varepsilon(\cdot)$ on $[\tau_1^\varepsilon, \tau_2^\varepsilon]$ converges to a straight line. By (A5.2), this straight line hits the boundary of the next decision set at precisely the same point where the $\bar{X}_\delta(\cdot)$ would hit it, if both started at $X_\delta(\tau_1)$ on $\partial G(1)$. A continuation of this argument yields that both the "diffusion" and "control" sections of $X_\delta(\cdot)$ are those of $\bar{X}_\delta(\cdot)$, which is what was to be proved. □

6. A numerical method for approximating the optimal value function and control. The control problem defined by the cost (4.8), system (3.14) and the control actions described by the possibilities associated with the off/on impulses associated with the discussion about Fig. 4.1 can be approximated by the numerical methods studied in [9]. The method in [9] involves a Markov chain (indexed by a "finite difference" parameter) approximation to the optimal continuous-time problem. We then show that the sequence of value functions for the chains converges to the optimal value function for the continuous parameter problem, and that suitable continuous parameter interpolations of the chain converge weakly to the optimal controlled continuous parameter process. The methods of [9] can be readily adapted to our problem, and only an outline will be given. The weak convergence methods used in [9] will have to be replaced by the methods here—owing to the reflection term, but the general idea is the same.

Let h be a finite-difference parameter, and let B_i be integral multiples of h. Let G_h denote the h-grid on $G = [0, B_1] \times [0, B_2]$. Define a_{ij} by $\Sigma_{ij}(t) = \int_0^t a_{ij}(X(s))\, ds$, and omit the x-argument in the $a_{ij}(\cdot)$ and $b^i(\cdot)$ below. For the Markov chain approximation, the status of the controls at any time is defined by the vector $P = (P^{01}, P^{02}, P^1, P^{12})$, where $P^\alpha = 1$ (respectively, 0) denotes that the control is on (the link is operating normally) (respectively, closed). Recall that, when $P = (1, 1, 1, 1)$, we write $P = 1$.

Let $\{X_n^h\}$ denote the approximating Markov chain, and let x denote the canonical current state, y the canonical successor state, and P_1 the canonical control that will be used at state x to bring the chain to the next state. Define $X^h(\cdot)$, the interpolated process, to be the right continuous piecewise constant process with interpolation intervals $\Delta t^h(x, P_1)$. Both these intervals and the transition probabilities $p^h(x, y \mid P_1)$ depend on the new chosen control as well as on the current state. If $P_1 \neq 1$, we use $\Delta t^h(x, P_1) = 0$; i.e., the interpolation interval has zero length. In this case, several steps of $\{X_n^h\}$ all occur simultaneously in the interpolation $X^h(\cdot)$. Define $Q_h(x) = 2[a_{11} + a_{22} - |a_{12}|] + h(|b^1| + |b^2|)$, and let $a_{ii} - |a_{12}| \geqq 0$, $i = 1, 2$. For $P_1 = 1$, we use $\Delta t^h(x, P_1) = h^2 / Q_h(x)$.

We now define the transition probabilities $p^h(x, y \mid P_1)$ for the chain when $P_1 = 1$, for $x, y \in G_h$. Let e_i denote the unit vector in the ith coordinate direction. We use

$$
\begin{aligned}
&p^h(x, x \pm e_i h \mid P_1 = 1) = [a_{ii} - |a_{12}| + h(b^i)^\pm] / Q_h(x), \\
&p^h(x, x + e_1 h - e_2 h \mid P_1 = 1) = p^h(x, x - e_1 h + e_2 h \mid P_1 = 1) \\
&\qquad = |a_{12}| / Q_h(x).
\end{aligned} \tag{6.1}
$$

If some x^i (the ith component of x) equals zero—then the transition probability (6.1) is modified as follows, as a concatenation of two transitions, the first being (6.1). For the second (the "reflection") step, we distinguish two cases.

Case 1. Either ($y^1 \geqq 0$, $y^2 < 0$) or ($y^1 < 0$, $y^2 \leqq 0$). Then simply project (reflect) the process back to the nearest point in G_h.

Case 2. $y^1<0$, $y^2>0$. Project to $(0, y^2)$ with probability $1-p_{12}/(p_{12}+p_{10})$ and to $(0, y^2-h)$ with probability $p_{12}/(p_{12}+p_{10})$. This step is to account for the $p_{12}y^1/(p_{12}+p_{10})$ term in (3.14).

Let P denote the control used to get the current state x. The actual state for the problem is the pair (x, P), since the cost associated with the next transition depends on whether or not some element of the current control vector is changed. Let $K^h(x, P, P_1)$ denote the costs associated with the transition, when current state is x, and control P changes to P_1. For $P_1=1$, $K^h(x, P, 1)=\Delta t^h(x, 1)k(x)$, the holding cost only.

We now define some of the transition probabilities and costs when $P_1\neq 1$. There are 15 possibilities, and only some typical ones will be described. These are constructed so that the limit (as $h\to 0$) of $X^h(\cdot)$ will be the reflected controlled $X(\cdot)$, and so that the associated costs for $X^h(\cdot)$ will also converge to that for $X(\cdot)$. We write $P=(P^{01}, P^{02}, P^{12}, P^1)$, $P_1=(P_1^{01}, P_1^{02}, P_1^{12}, P_1^1)$.

Let $P_1^{01}=0$, with other $P_1^\alpha=1$. Then use $p^h(x, x-e_1h\,|\,P_1)=1$ (by (A2.1), $x^1>0$ here) and $K^h(x, P, P_1)=q_{01}h+k_{01}I_{\{p^{01}=1, p_1^{01}=0\}}$. Now, let $P_1^{02}=0$ with other $P_1^\alpha=1$. Then $p^h(x, x-e_2h\,|\,P_1)=1$ and $K^h(x, P, P_1)=q_{02}h+k_{02}I_{\{p^{02}=1, p_1^{02}=0\}}$. For $P_1^{12}=0$ and other $P_1^\alpha=1$, we have $p^h(x, x-e_2h\,|\,P_1)=1$ and $K^h(x, P, P_1)=q_{12}h+k_{12}I_{\{p^{12}=1, p_1^{12}=0\}}$.

Now, let $P_1^1=0$ with other $P_1^\alpha=1$. Let $p_{12}g_{d1}\leqq g_{a1}$ (the reverse case is treated analogously) and refer to Fig. 6.1. The line from x to (a) is the mean direction of the appropriate impulse, and its slope (see § 4) is $[g_{a2}-(1-p_{22})g_{d2}]/g_{a1}=-p_{12}g_{d1}/g_{a1}$.

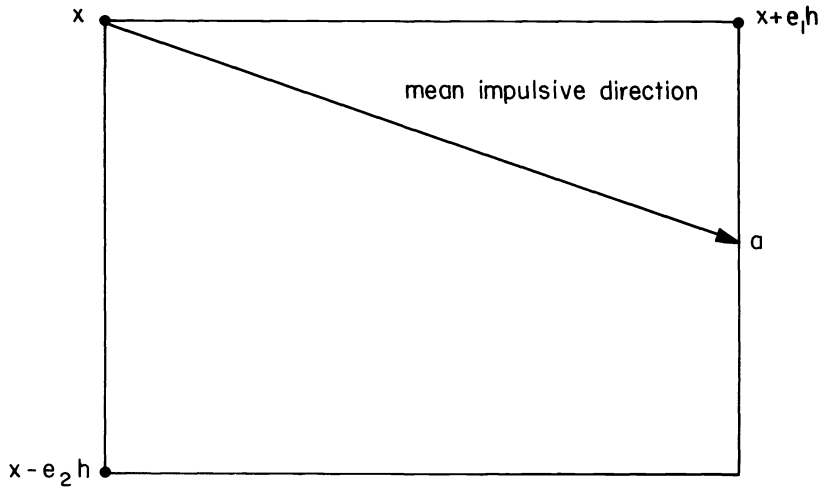

FIG. 6.1. $P^h(x, x+e_1h-e_2h/P_1)=1-P^h(x, x+e_1h/P_1)=p_{12}g_{d1}/g_{a1}$, $p_{12}g_{d1}\leqq g_{a1}$. *Transition probabilities for* $P_1^1=0$, *other* $P_1^\alpha=1$.

To "simulate" this mean line, we use

$$p^h(x, x+e_1h-e_2h\,|\,P_1)=p_{12}g_{d1}/g_{a1}=1-p^h(x, x+e_1h\,|\,P_1).$$

The instantaneous cost is $K^h(x, P, P_1)=k_1I_{\{p^1=1, p_1^1=0\}}$.

Now, let $P_1^{12}=P_1^{02}=0$ with all other $P_1^\alpha=1$. Then $p^h(x, x-e_2h\,|\,P_1)=1$. The "impulsive" part of $K^h(x, P, P_1)$ is obvious, namely, $k_{12}I_{\{p^{12}=1, p_1^{12}=0\}}+k_{02}I_{\{p^{02}=1, p_1^{02}=0\}}$. But the "opportunity" cost—that due to Z^{12} and U^{02} is less obvious. This is obtained from the relative rates at which $X^2(\cdot)$ decreases due to the effects of P_{12} and P_{02}

(respectively) being off. This is (respectively) $p_{12}g_{d1}$ and g_{a2}. Thus we use the "opportunity" cost

$$h[q_{12}p_{12}g_{d1}+q_{02}g_{a2}]/(p_{12}g_{d1}+g_{a2}).$$

The $p^h(x, y \mid P_1)$ and $K^h(x, P, P_1)$ are calculated in a similar way for all the other possibilities.

The dynamic programming equation for our "approximation" problem is

$$V^h(x, P)=\min_{P_1}\left[(\exp -\beta\Delta t^h(x, P_1))\sum_y p^h(x, y \mid P_1)V^h(y, P_1)+K^h(x, P, P_1)\right]. \tag{6.2}$$

The weak convergence methods of this paper can be used to show that $V^h(x, P)\to V(x, P)=\inf_{\pi\,\text{adm.}} V(\pi, x, P)$. For reasonable grid sizes, say 50×50, the numerical problem is quite tractable.

For the numerical problem, we do not need to duplicate the dynamics of the original system $X^\varepsilon(\cdot)$, but we can use any controlled process having the same controlled limit equation. See [9] for a fuller development of this computational point of view for a large class of more classical problems.

Appendix. We have used the assumption that if the first queue is empty at the start of a new "service interval" of length Δ, and an arrival occurs at some $\Delta'<\Delta$ later, then the service interval for that arrival is the residual time $\Delta-\Delta'$, and similarly for the second queue. In Theorem 1 of [11, p. 159], Iglehart and Whitt have shown that such an assumption does not affect the limit equations. A proof very similar to theirs works here, and we only outline the ideas and differences—with heavy reference to the cited theorem. We work only with the first queue because the second is treated in essentially the same way.

Let $\tilde{Q}^{1,\varepsilon}(\cdot)$ and $\tilde{X}^{1,\varepsilon}(\cdot)$ (set $\tilde{Q}^{1,\varepsilon}(\cdot)=\tilde{X}^{1,\varepsilon}(\cdot)/\sqrt{\varepsilon}$, $Q^{1,\varepsilon}(\cdot)=X^{1,\varepsilon}(\cdot)/\sqrt{\varepsilon}$) denote the quantities that would be obtained for the true queue; i.e., where an arrival to an empty queue has the correct—not the residual—service time. In [11], a sequence of potential "service times" has been constructed that (or some subsequence of which) has been used for *both* $Q^{1,\varepsilon}(\cdot)$ and $\tilde{Q}^{1,\varepsilon}(\cdot)$, and this has enabled a comparison of the two processes. We do the same thing here, following the method of [11] for their case $s=1$. The only difference is due to the state-dependence of the intervals here. Except for the t_0 of [11, p. 160], we use our own terminology (our $(Q, \tilde{Q}, \varepsilon, \Delta_k^{1,\varepsilon})$ is their $(Q', Q, 1/n, v_k^1)$). (First note an addendum to the proof in [11], for a case omitted there. If, given t, $\tilde{Q}^{1,\varepsilon}(\tau)\geqq 1$ for all $\tau\leqq t$, then set their $t_0=0$ and use $\tilde{Q}^{1,\varepsilon}(0)=Q^{1,\varepsilon}(0)$, and proceed as in their construction.) Let $Q^\varepsilon(0)\geqq 1$; otherwise, the result can be deduced from the argument below. The actual sequence of service intervals will be the same for both $\tilde{Q}^1$ and Q^1 until the first time that they equal zero.

Suppose that $\tilde{Q}^{1,\varepsilon}(\cdot)$ reaches zero at time τ, at the end of the kth service interval. Then generate a service interval $\Delta_{k+1}^{1,\varepsilon}$ as in the text, but with distribution determined by $\tilde{X}^\varepsilon(\tau)$ (the first component of which equals zero). Suppose that no arrival occurs between τ and $\tau+\Delta_{k+1}^{1,\varepsilon}$. Then generate the next interval $\Delta_{k+2}^{1,\varepsilon}$ as in the text, but with the distribution being determined by $\tilde{X}^\varepsilon(\tau+\Delta_{k+1}^{1,\varepsilon})$. On the other hand, suppose that an actual arrival occurs at $\tau<\tau'<\tau+\Delta_{k+1}^{1,\varepsilon}$. Then for $\tilde{Q}^{1,\varepsilon}(\cdot)$ generate another potential service interval $\Delta_{k+2}^{1,\varepsilon}$, with the distribution determined by $\tilde{X}^\varepsilon(\tau')$. We continue in this way to generate the (used and unused) service intervals for $\tilde{Q}^{1,\varepsilon}(\cdot)$. This sequence has the correct distribution for the true queue. Use the sequence $\{\Delta_k^{1,\varepsilon}\}$ for the queue $Q^{1,\varepsilon}(\cdot)$, with "residual" times used as in the text.

This procedure is precisely the one used in Theorem 1 of [11] except there the intervals are all independently and identically distributed. But, the association of particular intervals with particular service periods for both queues is exactly the same. We need only show that (A2.2), (A2.3), (A2.6) hold. In Theorem 1 of [11] it has been shown that for any T (our notation) $\sup_{t\leq T}|\tilde{X}^{1,\varepsilon}(t)-X^{1,\varepsilon}(t)|\to 0$ if (our notation) $\sqrt{\varepsilon}$ (maximum length of generated service intervals unused by $\tilde{Q}^{1,\varepsilon}(\cdot)$ on $[0, T]$) $\overset{\varepsilon}{\to} 0$. But this holds under our construction and the conditions on the intervals. This result, and the continuity and boundedness of $d_1(\cdot)$ yield

$$\begin{aligned} E^{1,\varepsilon}_{d,n}\Delta^{1,\varepsilon}_{n+1} &= g_{d1}+\sqrt{\varepsilon}\, d_1(\tilde{X}^{\varepsilon}_{S^{1,\varepsilon}_{d,n}})+o(\sqrt{\varepsilon}) \\ &= g_{d1}+\sqrt{\varepsilon}\, d_1(X^{\varepsilon}_{S^{1,\varepsilon}_{d,n}})+o(\sqrt{\varepsilon}), \end{aligned}$$

and similarly for the expression for the conditional variance. Thus, the generated service time sequence satisfies (A2.2), (A2.3), and (A2.6).

REFERENCES

[1] M. I. REIMANN, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[2] J. M. HARRISON, *Brownian models of queueing networks with heterogeneous customer populations*, Stanford Graduate School of Business Report, Stanford, CA, September 1986; in Proc. Inst. of Math. and Applications, Proc. on Stochastic Differential Equations, Springer-Verlag, Berlin, New York, 1987.

[3] ———, *The diffusion approximation for tandem queues in heavy traffic*, Adv. Appl. Probab., 10 (1978), pp. 886–905.

[4] A. J. LEMOINE, *Networks of queues—a survey of weak convergence results*, Management Sci., 24 (1978), pp. 1175–1193.

[5] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes: with Application to Stochastic Systems Theory*, M.I.T. Press, Cambridge, MA, 1984.

[6] H. J. KUSHNER AND W. RUNGGALDIER, *Filtering and control for wide bandwidth noise driven systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 123–133.

[7] H. J. KUSHNER AND K. M. RAMACHANDRAN, *Nearly optimal singular controls for wideband noise driven systems*, LCDS Report 86-43, Brown University, Providence, RI; SIAM J. Control Optim., 26 (1988), pp. 569–591.

[8] H. KUSHNER AND W. RUNGGALDIER, *Nearly optimal state feedback controls for stochastic systems with wideband noise disturbances*, SIAM J. Control Optim., 25 (1987), pp. 289–315.

[9] H. J. KUSHNER, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[10] J. M. HARRISON, T. M. SELLKE, AND A. J. TAYLOR, *Impulse control of Brownian motion*, Math. Oper. Res., 8 (1983), pp. 454–466.

[11] D. L. IGLEHART AND W. WHITT, *Multiple channel queues in heavy traffic.* I, Adv. in Appl. Math., 2 (1970), pp. 150–177.

[12] J. M. HARRISON AND M. I. REIMANN, *Reflected Brownian motion on an orthant*, Ann. Probab., 9 (1981), pp. 302–308.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1319-1332, November 1989

005

ON DISTURBANCE DECOUPLING IN DESCRIPTOR SYSTEMS*

L. R. FLETCHER† AND A. AASARAAI†

Abstract. The disturbance decoupling problem for a linear multivariable control system is to determine when and how a state feedback may be chosen so that some class of disturbances has no effect on the system output. The development of necessary and sufficient conditions for the solvability of this problem for state space systems by Wonham and his co-workers has profoundly influenced geometric control theory, particularly by means of the concept of an (A, B)-invariant subspace. This paper develops necessary and sufficient conditions for solvability of this problem for linear systems described by a mixture of algebraic and differential equations. Although the argument presented has the same structure as in the state space case, and incorporates a generalisation of (A, B)-invariant subspaces to descriptor systems, there are some additional issues. Demanding particularly careful attention is the ensuring of existence and uniqueness of classical solutions of the underlying differential equations, which cannot be taken for granted. The geometric structure of the solution space is investigated and a general notion of invariance of subspaces for descriptor systems is proposed.

Key words. descriptor systems, smooth solutions, disturbance decoupling, invariant subspaces

1. Introduction.

We study a linear multivariable control system in descriptor form:

$$E\dot{x}(t) = Ax(t) + Bu(t) + Sd(t), \qquad x(0) = x_0, \tag{1}$$

$$y(t) = Cx(t). \tag{2}$$

Here x, y, u, d are functions of time with values in $\mathscr{X} = \mathbb{R}^n$, $\mathbb{R}^p$, $\mathbb{R}^m$, $\mathbb{R}^q$, respectively, and E, A, B, C, S are real, constant, matrices of suitable sizes. Although we have in mind that E is singular, this is nowhere essential to our argument. Furthermore, many papers on such systems assume that E and A are square; following Wong [8], we will not make this assumption. A selection of examples of systems where descriptor models are useful is given by Luenberger [4] and a recent survey of results has been provided by Lewis [3]; however, implicit in this latter paper is the assumption that E and A are square. There are also intriguing, but unexplored, similarities to the work of Manitius [5] on retarded differential systems.

In (1) the function d represents unknown or unmodeled disturbances affecting the system. We investigate the questions of when and how the real $m \times n$ matrix F in the state feedback control law

$$u(t) = -Fx(t) \tag{3}$$

may be chosen so that the output $y(t)$ of the closed loop system (1)-(3) is independent of $d(t)$. In the state space case, that is when E is a nonsingular square matrix, this problem is discussed in detail in Chapter 4 of [9] and a complete solution is presented, using constructive arguments to obtain conditions on the system necessary and sufficient for the existence of F. Specifically, the following is a simple extension of Theorem 4.2 of [9]; here and throughout this paper we use $\mathscr{B}$ to denote the subspace spanned by the columns of the matrix B.

THEOREM 1.1. *Suppose E is a nonsingular square matrix. Then a necessary and sufficient condition for the existence of a feedback F such that for any continuous function d there exists an initial condition x_0 for which $y(t) \equiv 0$ is the existence of a subspace $\mathscr{V}$*

* Received by the editors October 19, 1987; accepted for publication (in revised form) December 23, 1988.
† Department of Mathematics and Computer Science, University of Salford, Salford, Lancashire M5 4WT, United Kingdom.

of ker C *such that*

$$A\mathcal{V} \subseteq E\mathcal{V} + \mathcal{B}, \qquad \text{Im } S \subseteq E\mathcal{V}. \tag{4}$$

It is a straightforward deduction from the proof of this result given by Wonham [9] that, for some suitably chosen F, a given initial condition x_0 leads to $y(t) \equiv 0$ whatever the disturbance function $d(t)$ if and only if $x_0 \in \mathcal{V}^*$, the unique maximal element of the set of subspaces $\mathcal{V}$ satisfying the conditions in the theorem.

Our aim in this paper is the generalisation of Theorem 1.1 to the case when E is singular and perhaps not square. Some work on disturbance decoupling in descriptor systems appears in [10] but there the class of feedbacks considered is different from ours as defined in (3). Our arguments follow closely those in [9] but there is a major new issue to be attended to in the descriptor case. It is essential that the feedback F be chosen so that the closed loop system

$$E\dot{x} = (A - BF)x + Sd, \qquad x(0) = x_0 \tag{5}$$

has appropriate regularity properties. To be precise about our objectives in this respect we need to refer to a time-invariant, first-order, linear system of differential and algebraic equations

$$E\dot{x}(t) = Mx(t) + f(t), \qquad x(0) = x_0 \tag{6}$$

in which E and x are as before, M is a matrix of the same size as E, and $f(t)$ is a vector function of time with the appropriate number of components. The theory of systems such as (5) or (6) depends on what is meant by a solution; throughout this paper we adopt the following definition.

DEFINITION 1.1. A function $x : [0, \infty) \to \mathbb{R}^n$ is a *solution* of the initial value problem (6) if

(i) $\dot{x}(t)$ exists and satisfies the differential equation for all $t > 0$;

(ii) $x(0) = x_0$;

(iii) $\dot{x}(0+)$ exists and $E\dot{x}(0+) = Mx_0 + f(0)$.

This stipulation about the admissible solutions to (5) or (6) implies some regularity properties of the function $f(t)$ and an appropriate relation between the behaviour of f at zero and x_0. It is recognised that this is suitable for some applications and not for others [6]; however, it is essential in all approaches that F be chosen so that uniqueness of solutions to (5) prevails. Thus we have the following definition.

DEFINITION 1.2. We shall say that the system (6) is *semiregular* if it has at most one solution for any initial condition.

It is shown in [1] that (6) is semiregular if and only if the matrix $M - \lambda E$ has linearly independent columns for some value of λ. Moreover, this is equivalent to requiring that the pencil $M - \lambda E$ is nonsingular in the sense of [8] so this definition is a generalisation for possibly nonsquare E and A of the classical notion of regularity [3, § 2]. Ultimately, a feedback F that achieves disturbance decoupling must also ensure that the system (5) is semiregular in this sense, so we need to avoid at the outset the pathological situation that for no F does closed loop regularity hold. It is shown in [1] that this is equivalent to the following assumption.

REGULARISABILITY ASSUMPTION. We will assume that for at least one complex number λ the matrix

$$[A - \lambda E, B]$$

has (at least) n linearly independent columns.

Note that we do not need to assume that the open loop system enjoys uniqueness of solutions. On the other hand our regularisability assumption can only be satisfied if in (1) there are at least n equations.

In discussing (6) it is important to bear in mind that, eventually, in constructing a feedback matrix to solve the disturbance decoupling problem, the matrix M is replaced by the closed loop system matrix $A-BF$, which is unknown at the outset, and $f(t)$ represents the disturbance function $Sd(t)$ where $d(t)$ is unknown throughout. For these reasons the well-known analysis of (6) by means of the Kronecker canonical form (see, for example, [7]) does not meet our needs. Although it might be possible to describe by such means the features of the pencil $A-\lambda E$ that are desirable for disturbance decoupling, it is not known at present to what extent the Kronecker canonical form of $A-BF-\lambda E$ can be assigned by a suitable choice of the matrix F. We must therefore adopt a different approach to (4) as exemplified by the following key concept.

DEFINITION 1.3. A pair of subspaces $\mathcal{U}$, $\mathcal{V}$ will be said to be $(\{A, E\}, B)$-*invariant* if

(i) $A\mathcal{V} \subseteq E\mathcal{V} + \mathcal{B}$;

(ii) $\mathcal{V} \cap \ker E = \{0\}$;

(iii) $E\mathcal{U} \subseteq A\mathcal{U} + \mathcal{B}$;

(iv) $\dim (E\mathcal{U} \cap \mathcal{B}) \leqq \dim \{u \in \mathcal{U}: Au \in \mathcal{B}\}$.

The final ingredient in our main result is a statement of what we mean by "disturbance decoupling."

DEFINITION 1.4. We shall say that a feedback matrix F *achieves disturbance decoupling* for the system (1), (2) if the closed loop system (5), (2) has the following properties.

(i) For no initial condition x_0 and disturbance function $d(t)$ does (5) have more than one solution;

(ii) For every disturbance function $d(t)$ that is sufficiently differentiable there is a solution of (5) for any initial value x_0 in a certain affine subspace of $\mathcal{X}$;

(iii) $y(t)$ is independent of $d(t)$ in that for any sufficiently differentiable d there are initial conditions, admissible in the sense of (ii), such that $y(t) \equiv 0$.

Now we can state our main result.

MAIN THEOREM. *So that there exists a matrix F achieving disturbance decoupling for the system* (1), (2) *it is necessary and sufficient that there exists an* $(\{A, E\}, B)$-*invariant pair of subspaces $\mathcal{U}$, $\mathcal{V}$ contained in* $\ker C$ *such that*

$$\operatorname{Im} S \subseteq E\mathcal{V} + \mathcal{L}$$

for some $\mathcal{L}$ satisfying the following:

(a) $E\mathcal{U} \subseteq \mathcal{L} \subseteq A\mathcal{U} + \mathcal{B}$;

(b) $\dim \mathcal{L} \cap \mathcal{B} \leqq \dim \{u \in \mathcal{U}: Au \in \mathcal{B}\}$.

We have not been able to determine simple conditions on x_0 necessary and sufficient for the existence of a solution (in the sense of Definition 1.1) of (5) for a suitably chosen F in which $y(t) \equiv 0$ whatever the disturbance function $d(t)$. The main obstacle is that the admissibility of x_0 depends on the values of d and its derivatives at $t=0$ so that a given initial condition, $x_0 = 0$ say, might be admissible for one disturbance function and not for another; this issue is addressed in (iii) of Definition 1.4. In the proof of the Main Theorem the value of $v(0)$ in (19) below is the crucial point. It is not difficult to deduce some sufficient conditions on x_0 from the argument at the end of § 3. For example, if the disturbance function $d(t)$ is such that $Sd(t) \in E\mathcal{V}$ in some interval $[0, t_0)$ and $x_0 \in \mathcal{V}$ then, for some suitably chosen F, $y(t) \equiv 0$. However, these conditions are not necessary.

The linear algebraic arguments in § 2 below form the heart of the paper. They provide a feedback characterisation of those subspaces that are invariant in the sense

of satisfying Definition 1.3. In § 3 we prove the sufficiency claim in our Main Theorem. Finally, in § 4, we examine the Kronecker canonical form of a closed loop system in which disturbance decoupling is assumed to have been achieved, in order to prove the necessity part of the Main Theorem.

2. Invariant subspaces for descriptor systems. Our main objective in this section is a proof of the following key result:

THEOREM 2.1. *If the pair of subspaces $\mathcal{U}$, $\mathcal{V}$ is $(\{A, E\}, B)$-invariant, then there exists a matrix F and a subspace $\mathcal{W}$ with $\mathcal{V} \subseteq \mathcal{W} \subseteq \mathcal{U} + \mathcal{V}$ such that*

$$(A - BF)\mathcal{W} \subseteq E\mathcal{W}, \qquad E\mathcal{U} \subseteq (A - BF)\mathcal{U}$$

and the matrix $A - BF - \lambda E$ has linearly independent columns for some complex number λ.

Our proof of this result is couched entirely in terms of linear algebra and is somewhat intricate so we proceed by means of a sequence of lemmas. The first of these is of some interest in its own right.

LEMMA 2.2. *For a given subspace $\mathcal{U}$ there exists a matrix F such that*

$$E\mathcal{U} \subseteq (A - BF)\mathcal{U} \tag{7}$$

if and only if $\mathcal{U}$ satisfies (iii) *and* (iv) *of Definition* 1.3. *Moreover, in these circumstances we can arrange that $\mathcal{L} \subseteq (A - BF)\mathcal{U}$ for any subspace $\mathcal{L}$ satisfying conditions* (a) *and* (b) *in the Main Theorem.*

Proof. Let $u_1, \cdots, u_s$ be elements of $\mathcal{U}$ such that $Eu_1, \cdots, Eu_r$ is a basis of $E\mathcal{U} \cap \mathcal{B}$ and $u_{r+1}, \cdots, u_s$ is a basis of $\mathcal{U} \cap \ker E$ and let $u_{s+1}, \cdots, u_t$ be the remainder of a basis of $\mathcal{U}$.

Suppose (7) holds for some matrix F then, clearly, (iii) of Definition 1.3 holds. Furthermore, there exist $w_1, \cdots, w_r \in \mathcal{U}$ such that

$$Eu_i = (A - BF)w_i \quad \text{for } i = 1, \cdots, r$$

and so for $i = 1, \cdots, r$ we have

$$Aw_i = Eu_i + BFw_i \in \mathcal{B}.$$

Thus $w_1, \cdots, w_r$ are linearly independent elements of the subspace $\{u \in \mathcal{U}: Au \in \mathcal{B}\}$ so (iv) of Definition 1.3 holds.

Conversely, suppose (iii) and (iv) of Definition 1.3 hold. Then

$$Eu_i = Aw_i + Bv_i \quad \text{for } i = 1, \cdots, t \tag{8}$$

for some $w_1, \cdots, w_t \in \mathcal{U}$. Clearly, $w_{s+1}, \cdots, w_t$ are linearly independent and we can take $w_1 = \cdots = w_r = 0$. We are assuming that there exist linearly independent vectors $\tilde{w}_1, \cdots, \tilde{w}_r \in \mathcal{U}$ such that

$$A\tilde{w}_i = B\tilde{v}_i \quad \text{for } i = 1, \cdots, r.$$

We show first that the vectors $\tilde{w}_1, \cdots, \tilde{w}_r$; $w_{s+1}, \cdots, w_t$ are linearly independent. Indeed were there scalars $\alpha_{s+1}, \cdots, \alpha_t$, not all zero, such that

$$\sum_{i=s+1}^{t} \alpha_i Aw_i \in \mathcal{B},$$

then, according to (8), the vector

$$\sum_{i=s+1}^{t} \alpha_i Eu_i = \sum_{i=s+1}^{t} \alpha_i Aw_i + B \sum_{i=s+1}^{t} \alpha_i v_i$$

would also lie in $\mathcal{B}$, contrary to the definition of the integer s. Now define F by

$$F\tilde{w}_i = \tilde{v}_i - v_i, \qquad i = 1, \cdots, r,$$

$$Fw_i = -v_i, \qquad i = s+1, \cdots, t$$

and extend F to the whole of $\mathcal{X}$ arbitrarily. Then for $i = 1, \cdots, r$ we have

$$\begin{aligned}(A - BF)\tilde{w}_i &= B\tilde{v}_i - B(\tilde{v}_i - v_i) \\ &= Bv_i = Eu_i\end{aligned}$$

for $i = r+1, \cdots, s$ we have

$$Eu_i = (A - BF)0$$

and for $i = s+1, \cdots, t$ we have

$$(A - BF)w_i = Aw_i + Bv_i = Eu_i.$$

Thus $E\mathcal{U} \subseteq (A - BF)\mathcal{U}$ as required.

Now suppose $\mathcal{Z}$ is a subspace satisfying conditions (a) and (b) of the Main Theorem. Let $z_1, \cdots, z_p$ be a basis of $\mathcal{Z} \cap \mathcal{B}$. Then, by condition (b), there exist linearly independent elements $v_1, \cdots, v_p$ of $\mathcal{U}$ such that $Av_i \in \mathcal{B}$. There exists a matrix F such that (7) holds, so there are vectors $v_1, \cdots, v_p \in \mathcal{U}$ and $g_1, \cdots, g_p$ satisfying

$$z_i = (A - BF)v_i + Bg_i \quad \text{for } i = 1, \cdots, p. \tag{9}$$

Now let $z_{p+1}, \cdots, z_q$ be linearly independent elements of $E\mathcal{U}$ such that $z_1, \cdots, z_q$ is a basis of $(\mathcal{Z} \cap \mathcal{B}) + E\mathcal{U}$. Then for $i = p+1, \cdots, q$ there exist $u_i, v_i \in \mathcal{U}$ such that

$$Eu_i = z_i = (A - BF)v_i. \tag{10}$$

Our next step is to show that the subspaces of $\mathcal{U}$ spanned by $v_1, \cdots, v_p$ and by $v_{p+1}, \cdots, v_q$ intersect trivially. First note that, since $E\mathcal{U} \cap \mathcal{B} \subseteq \mathcal{Z} \cap \mathcal{B}$, the space spanned by $z_{p+1}, \cdots, z_q$ does not intersect $\mathcal{B}$. Now suppose

$$\sum_{i=1}^{p} \alpha_i v_i = \sum_{i=p+1}^{q} \beta_i v_i.$$

Then, by (9) and (10)

$$\sum_{i=1}^{p} \alpha_i z_i - B \sum_{i=1}^{p} \alpha_i g_i = \sum_{i=p+1}^{q} \beta_i z_i.$$

But the left-hand side here lies in $\mathcal{B}$, and, as we have just remarked, the right-hand side does not, unless it is zero. Thus the subspaces of $\mathcal{U}$ we are considering do intersect trivially so there exists a matrix F_0 such that

$$F_0 v_i = -g_i, \qquad i = 1, \cdots, p,$$

$$F_0 v_i = 0, \qquad i = p+1, \cdots, q.$$

Putting $F_1 = F + F_0$ we see that

$$z_i = (A - BF_1)v_i \quad \text{for } i = 1, \cdots, q$$

and

$$E\mathcal{U} \subseteq (A - BF_1)\mathcal{U}.$$

To complete the proof of Lemma 2.2, let $z_{q+1}, \cdots, z_r$ be the remainder of a basis of $\mathscr{Z}$ and choose a matrix F_1 so that for the largest possible integer k there exist $v_1, \cdots, v_k \in \mathscr{U}$ such that

$$z_i = (A - BF_1)v_i \quad \text{for } i = 1, \cdots, k. \tag{11}$$

To complete the proof of this lemma we show that $k = r$; we have shown in the previous paragraph that $k \geqq q$. Suppose $k < r$; then, by condition (a) of the Main Theorem, there is some nonzero g_{k+1} such that

$$z_{k+1} = (A - BF_1)v_{k+1} + Bg_{k+1}.$$

Were v_{k+1} dependent on $v_1, \cdots, v_k$ in (11), say

$$v_{k+1} = \sum_{i=1}^{k} \alpha_i v_i$$

then we would have

$$\sum_{i=1}^{k} \alpha_i z_i = z_{k+1} - Bg_{k+1}$$

so that

$$Bg_{k+1} = z_{k+1} - \sum_{i=1}^{k} \alpha_i z_i \in \mathscr{Z} \cap \mathscr{B}.$$

But $z_1, \cdots, z_p$ is a basis of $\mathscr{Z} \cap \mathscr{B}$ and $k+1 > p$ so we would have a contradiction. Since, then, v_{k+1} is not dependent on $v_1, \cdots, v_k$ there exists a matrix F_2 such that

$$F_2 v_i = 0 \quad \text{for } i = 1, \cdots, k,$$

$$F_2 v_{k+1} = -g_{k+1}.$$

Replacing F_1 by $F_1 + F_2$ would increase the value of k, which would be a contradiction. Thus $k = r$.

This completes the proof of Lemma 2.2.

It is worth noting that conditions (i) and (ii) of Definition 1.3 are necessary and sufficient for the existence of a matrix F such that

$$(A - BF)\mathscr{V} \subseteq E\mathscr{V} \tag{12}$$

and the matrix $A - BF - \lambda E$ has linearly independent columns for some complex number λ. As far as necessity is concerned, we need only point out that if (12) holds then

$$(A - BF - \lambda E)\mathscr{V} \subseteq E\mathscr{V}$$

so $A - BF - \lambda E$ having linearly independent columns implies that $\dim E\mathscr{V} = \dim \mathscr{V}$. Sufficiency will follow from Theorem 2.1 with $\mathscr{U} = \{0\}$.

Continuing with the proof of Theorem 2.1, from the set of F satisfying the requirements of Lemma 2.2, we aim to select one that also satisfies the other requirements of Theorem 2.1 for a suitable subspace $\mathscr{W}$.

LEMMA 2.3. *If $\mathscr{U}$ satisfies* (iii) *of Definition* 1.3 *then there exist subspaces $\mathscr{Q}$, $\mathscr{T}$ of $\mathscr{U}$ such that*

$$\mathscr{U} = \mathscr{Q} \oplus \mathscr{T}, \qquad E\mathscr{Q} \subseteq (A - BF)\mathscr{Q},$$

$$E\mathscr{T} = (A - BF)\mathscr{T}, \qquad \mathscr{T} \cap \ker E = \{0\}.$$

Proof. Let $G: \mathcal{U} \to \mathcal{U}$ be a linear mapping obtained by writing for $u \in \mathcal{U}$

$$Gu = w$$

where $w \in \mathcal{U}$ is a solution of the equation

$$Eu = (A - BF)w.$$

Now let

$$\mathcal{Q} = \bigcup_i \ker G^i, \qquad \mathcal{T} = \bigcap_i \operatorname{Im} G^i.$$

It is a well-known general property of linear transformations (see, for example, p. 48–49 of [2]) that, with these definitions

$$\mathcal{Q} \oplus \mathcal{T} = \mathcal{U}, \quad G\mathcal{Q} \subseteq \mathcal{Q}, \quad G\mathcal{T} = \mathcal{T}, \quad \mathcal{T} \cap \ker G = \{0\}.$$

It is easy to see that this completes the proof of Lemma 2.3.

DEFINITION 2.1. For an $(\{A, E\}, B)$-invariant pair $\mathcal{U}, \mathcal{V}$ let $\mathcal{W}$ be a subspace that is maximal subject to the following:

(i) $\mathcal{V} \subseteq \mathcal{W} \subseteq \mathcal{V} + \mathcal{T}$;

(ii) $A\mathcal{W} \subseteq E\mathcal{W} + \mathcal{B}$;

(iii) $\mathcal{W} \cap \ker E = \{0\}$ where $\mathcal{T}$ is the subspace of $\mathcal{U}$ whose existence is established in Lemma 2.3.

Since $\mathcal{V}$ itself satisfies these conditions such subspaces $\mathcal{W}$ exist although, in general, $\mathcal{W}$ is not uniquely determined. However, we have the following useful result.

LEMMA 2.4. *If $\mathcal{W}_1$, $\mathcal{W}_2$ are both maximal subject to satisfying* (ii) *and* (iii) *of Definition* 2.1, *then* $E\mathcal{W}_1 = E\mathcal{W}_2$.

Proof. We show that $\mathcal{W}_1 \subseteq \mathcal{W}_2 + \ker E$. Let $H: \mathcal{W}_1 \to \mathcal{W}_1$ be the linear transformation given by $Hw = z$ where $z \in \mathcal{W}_1$ is chosen to satisfy

$$Aw = Ez + b$$

for some $b \in \mathcal{B}$. If w_1 is an eigenvector of H, say $Hw_1 = \lambda w_1$, then, for some $b_1 \in \mathcal{B}$,

$$Aw_1 = \lambda Ew_1 + b_1.$$

Now the space spanned by w_1 and W_2 satisfies Definition 2.1(iii) and so, by the maximality of $\mathcal{W}_2$

$$w_1 \in \mathcal{W}_2 + \ker E.$$

Thus

$$Aw_1 = Ew_2 + b_1$$

for some $w_2 \in \mathcal{W}_2$.

Now suppose $v_1 \in \mathcal{W}_1$ such that $Hv_1 = \lambda v_1 + w_1$; that is, v_1 is a principal vector of H with respect to the eigenvalue λ. Then

$$\begin{aligned} Av_1 &= E(\lambda v_1 + w_1) + b_2 \\ &= \lambda Ev_1 + Ew_2 + b_2 \end{aligned}$$

for some $b_2 \in \mathcal{B}$. Now the subspace spanned by $\mathcal{W}_2$ and v_1 satisfies Definition 2.1(iii) so, by the maximality of $\mathcal{W}_2$

$$v_1 \in \mathcal{W}_2 + \ker E.$$

Continuing in this way shows that the entire root space [2] of H in $\mathcal{W}_1$ relating to λ is contained in $\mathcal{W}_2+\ker E$. Repeating this for each eigenvalue of H shows that $\mathcal{W}_1 \subseteq \mathcal{W}_2+\ker E$.

This completes the proof of Lemma 2.4.

COROLLARY 2.5. *In the notation of Definition* 2.1

$$E\mathcal{T} \subseteq E\mathcal{W}.$$

Proof. The subspace $\mathcal{T}$ satisfies (ii) and (iii) of Definition 2.1.

LEMMA 2.6. *For an* $(\{A, E\}, B)$*-invariant pair* $\mathcal{U}$, $\mathcal{V}$ *the matrix* F *can be chosen to satisfy*

$$(A-BF)\mathcal{W} \subseteq E\mathcal{W}$$

(*in addition to the requirements of Lemma* 2.2).

Proof. Let $w_1, \cdots, w_k$ be a basis of $\mathcal{W} \cap \mathcal{T}$ and $w_{k+1}, \cdots, w_l$ be the remainder of a basis of $\mathcal{W}$. Then, for $i=1, \cdots, k$ there exist $t_1, \cdots, t_k \in \mathcal{T}$ such that

$$(A-BF)w_i = Et_i$$

and so, by Corollary 2.5,

$$(A-BF)w_i \in E\mathcal{W} \quad \text{for } i=1, \cdots, k.$$

Now $w_{k+1}, \cdots, w_l \in \mathcal{W} \setminus \mathcal{U}$ by the construction of $\mathcal{W}$ and also there exist $v_i \in \mathcal{W}$ and g_i such that

$$Aw_i = Ev_i + Bg_i \quad \text{for } i=k+1, \cdots, l.$$

If we arrange that

$$Fw_i = g_i \quad \text{for } i=k+1, \cdots, l,$$

then $F|\mathcal{U}$ can remain unaltered and

$$(A-BF)w_i \in E\mathcal{W} \quad \text{for } i=k+1, \cdots, l.$$

This completes the proof of Lemma 2.6.

Finally in this section we show that a matrix F can be chosen from amongst those satisfying the requirements of Lemmas 2.2 and 2.6 and such that, for some complex number λ, the matrix $A-BF-\lambda E$ has linearly independent columns. To do this, let $w_1, \cdots, w_k$ continue to denote a basis of $\mathcal{W}$, let $u_{k+1}, \cdots, u_l$ be linearly independent vectors in $\mathcal{Q}$ such that $Eu_{k+1}, \cdots, Eu_l$ complement $E\mathcal{W}$ in $E(\mathcal{U}+\mathcal{V})$ and let $w_{k+1}, \cdots, w_l$ be vectors in $\mathcal{Q}$ such that $Eu_i = (A-BF)w_i$ for $i=k+1, \cdots, l$.

LEMMA 2.7. *Suppose* F *satisfies the requirements of Lemmas* 2.2 *and* 2.6. *For all but a finite number of values of* λ, *if* w *is a nonzero linear combination of* $w_1, \cdots, w_l$ *then* $(A-BF-\lambda E)w \neq 0$.

Proof. The choice of $w_{k+1}, \cdots, w_l$ ensures that the vectors $(A-BF)w_i = Eu_i$ for $i=1, \cdots, l$ are linearly independent. Indeed if

$$(A-BF)\sum_{i=1}^{k} \alpha_i w_i = (A-BF)\sum_{i=k+1}^{l} \alpha_i w_i,$$

then, since $(A-BF)\mathcal{W} \subseteq E\mathcal{W}$, there exists $v \in \mathcal{W}$ such that

$$Ev = \sum_{i=k+1}^{l} \alpha_i Eu_i$$

contrary to the choice of $u_{k+1}, \cdots, u_l$ unless $\alpha_{k+1} = \cdots = \alpha_l = 0$. Now, if P is a matrix whose columns are a basis of $(E\mathcal{W})^{\perp}$ and $Q_0 = [w_{k+1}, \cdots, w_l]$ is a matrix with the

columns shown, then we have just proved that the matrix $P^T(A-BF-\lambda E)Q_0$ has linearly independent columns when $\lambda=0$. Hence it has linearly independent columns for all but a finite number of values of λ or, in other words, for only a finite number of values of λ is there a vector w such that $Ew \in E(\mathcal{U}+\mathcal{V})\backslash E\mathcal{W}$ and $(A-BF-\lambda E)w \in E\mathcal{W}$.

Now let $H: \mathcal{W} \to \mathcal{W}$ denote the linear transformation given by $Hw=v$ for $v, w \in \mathcal{W}$ satisfying

$$(A-BF)w = Ev.$$

Then

$$\begin{aligned}(A-BF-\lambda E)w &= Ev-\lambda Ew\\ &= E(H-\lambda I)w.\end{aligned}$$

If λ is not an eigenvalue of H then, recalling that $\mathcal{W}\cap\ker E=\{0\}$, it is clear that $(A-BF-\lambda E)w=0$ for $w\in\mathcal{W}$ only if $w=0$.

If λ does not belong to either of these exceptional sets then the conclusion of Lemma 2.7 holds.

Proof of Theorem 2.1. Let $w_{l+1}, \cdots, w_n$ be the remainder of a basis of $\mathcal{X}$. Choose F from amongst those that satisfy the requirements of Lemma 2.6 and λ so that, for the largest possible value of the integer r, the vectors

$$(A-BF-\lambda E)w_i \quad \text{for } i=1,\cdots,r \tag{13}$$

are linearly independent. We must show that $r=n$; Lemma 2.7 shows that $r\geqq l$. Suppose $r<n$. By our regularisability assumption, the matrix $[A-\mu E, B]$ has at least n linearly independent columns for some μ and, hence, for almost all μ. We can assume, therefore, that the λ just chosen is not one of these exceptional values of μ and hence that the vectors

$$(A-BF-\lambda E)w_i \quad \text{for } i=r+1,\cdots,n$$

are linearly dependent on those in (13) and that there are vectors $g_{r+1}, \cdots, g_n$ such that

$$Bg_{r+1}, \cdots, Bg_n \tag{14}$$

are linearly independent of those in (13). Define $F_1: \mathcal{X} \to \text{span}\{g_{r+1}, \cdots, g_n\}$ by

$$F_1 w_i = 0 \quad \text{for } i=1,\cdots,r,$$

$$F_1 w_i = g_i \quad \text{for } i=r+1,\cdots,n.$$

Since $r\geqq l$, $F_1 w=0$ for all $w\in\mathcal{W}$ and

$$Eu_i = (A-B(F+F_1))w_i \quad \text{for } i=k+1,\cdots,l,$$

$F+F_1$ satisfies the requirements of Lemmas 2.2 and 2.6. On the other hand, the vectors

$$(A-B(F+F_1)-\lambda E)w_i \quad \text{for } i=1,\cdots,n$$

span the same space as the vectors in (13) and (14). This contradicts the choice of F and λ so $r=n$.

This completes the proof of Theorem 2.1.

3. The proof of the Main Theorem: sufficiency. In this section we show that the conditions in the Main Theorem are sufficient to ensure the existence of a feedback matrix F achieving disturbance decoupling in the sense we have defined it. More precisely, we prove that any matrix F satisfying the conclusions of Theorem 2.1 and

Lemma 2.2 with respect to the given $\mathcal{U}$, $\mathcal{V}$, $\mathcal{Z}$ will suffice. Note that the uniqueness of solutions to the closed loop equation

$$E\dot{x} = (A - BF)x + Sd(t), \qquad x(0) = x_0 \tag{15}$$

follows, by Theorem 1 of [1], from the matrix $A - BF - \lambda E$ having linearly independent columns for some complex number λ. Thus it is sufficient to consider existence of solutions; the argument we provide illuminates some more general issues so we postpone any consideration of $y(t)$ until the end of this section.

For most of this section we discuss the existence of solutions of the system (6), assuming that the pencil $M - \lambda E$ is nonsingular in the sense of Wong [8]. We have already noted that the Kronecker canonical form is not an appropriate technique and we will also require information about the subspaces in which a solution evolves in more detail than is provided by existing approaches. On the other hand, most of the linear algebra we require is available in [8], though it will be necessary for us to reformulate some of Wong's results. We begin with a simple result about the case $f(t) \equiv 0$.

LEMMA 3.1. *Let $\mathcal{X}_0$ be a subspace of $\mathcal{X}$ and let $\mathcal{V}^*$ denote the unique maximal element of the collection*

$$\{\mathcal{V} \subseteq \mathcal{X}_0 \colon M\mathcal{V} \subseteq E\mathcal{V}\} \tag{16}$$

of subspaces of $\mathcal{X}_0$. Then the equation

$$E\dot{x}(t) = Mx(t), \qquad x(0) = x_0 \tag{17}$$

has a solution such that $x(t) \in \mathcal{X}_0$ for all $t \geqq 0$ if and only if $x_0 \in \mathcal{V}^*$.

Proof. Suppose $x_0 \in \mathcal{V}^*$. Since $M\mathcal{V}^* \subseteq E\mathcal{V}^*$, there exists a linear transformation $\mathcal{X} \to \mathcal{X}$ with matrix L such that $L\mathcal{V}^* \subseteq \mathcal{V}^*$ and $Mv = ELv$ for all $v \in \mathcal{V}^*$. If $x_0 \in \mathcal{V}^*$ then $x(t) = e^{Lt}x_0$ satisfies (17) and has $x(t) \in \mathcal{X}_0$ for all $t \geqq 0$.

Conversely, suppose $x(t)$ satisfies (17) with $x(t) \in \mathcal{X}_0$ for all $t \geqq 0$. Let $\mathcal{V}$ be the smallest subspace of $\mathcal{X}_0$ such that $x(t) \in \mathcal{V}$ for all $t \geqq 0$; we show that $\mathcal{V}$ is one of the subspaces (16). If $v \in \mathcal{V}$ then $v = x(t_1) + x(t_2) + \cdots + x(t_k)$ for some $t_1, t_2, \cdots, t_k$ and so

$$Mv = E\dot{x}(t_1) + E\dot{x}(t_2) + \cdots + E\dot{x}(t_k).$$

Thus, to show that $M\mathcal{V} \subseteq E\mathcal{V}$, it is sufficient to prove that $\dot{x}(t) \in \mathcal{V}$ for all $t \geqq 0$. But

$$\dot{x}(t) = \lim_{\delta t \to 0} \frac{x(t + \delta t) - x(t)}{\delta t} \tag{18}$$

(with $\delta t > 0$ if $t = 0$) where, by definition, $x(t + \delta t)$, $x(t) \in \mathcal{V}$. Moreover, $\mathcal{V}$ is closed since $\mathcal{X}$ is finite-dimensional, so proceeding to the limit in (18) does not leave $\mathcal{V}$. Thus $x(t) \in \mathcal{V}^*$ for all $t \geqq 0$ so $x(0) \in \mathcal{V}^*$, as required.

This completes the proof of Lemma 3.1.

The main result of this section is Theorem 3.2.

THEOREM 3.2. *Suppose $\mathcal{U}$ and $\mathcal{W}$ are subspaces of $\mathcal{X}$ satisfying*

$$E\mathcal{U} \subseteq M\mathcal{U}, \qquad M\mathcal{W} \subseteq E\mathcal{W}.$$

Then (6) has a solution $x(t)$ if f is differentiable dim $\mathcal{U}$ *times and such that*

$$f(t) \in E\mathcal{W} + M\mathcal{U} \quad \text{for all } t \geqq 0$$

and x_0 belongs to an appropriate affine subspace of $\mathcal{U} + \mathcal{W}$. Furthermore, $x(t) \in \mathcal{U} + \mathcal{W}$ for all $t \geqq 0$.

The key step in the proof of this theorem is a result in linear algebra. Define subspaces $\mathcal{U}^*$, $\mathcal{W}^*$ of $\mathcal{U}+\mathcal{W}$ as follows:

$$\mathcal{U}^* = \sup\{\mathcal{U}_1 \subseteq \mathcal{U}+\mathcal{W}\colon E\mathcal{U}_1 \subseteq M\mathcal{U}_1\},$$

$$\mathcal{W}^* = \sup\{\mathcal{W}_1 \subseteq \mathcal{U}+\mathcal{W}\colon M\mathcal{W}_1 \subseteq E\mathcal{W}_1\}.$$

Then we have Lemma 3.3.

LEMMA 3.3. *There exists a subspace $\mathcal{Q}$ of $\mathcal{X}$ such that*

$$E\mathcal{Q} \subseteq M\mathcal{Q}, \qquad \mathcal{U}+\mathcal{W} = \mathcal{W}^* \oplus \mathcal{Q}.$$

Proof. It follows from Lemma 3.1 that $\mathcal{W}^*$ is the set of admissible initial conditions for (10) which lie in $\mathcal{U}+\mathcal{W}$, analogous to the subspace H_I in [8]. Note that whatever the value of λ, $(M-\lambda E)\mathcal{W}^* \subseteq E\mathcal{W}^*$, so by the semiregularity of the pencil $M-\lambda E$, $\ker E \cap \mathcal{W}^* = \{0\}$ and, similarly, $\ker M \cap \mathcal{U}^* = \{0\}$. Furthermore, if $Mu = Ev$ for $u \in \mathcal{U}^*$, $v \in \mathcal{W}^*$ then, by maximality, $u \in \mathcal{W}^*$ so that $E\mathcal{W}^* \cap M\mathcal{U}^* = M(\mathcal{U}^* \cap \mathcal{W}^*)$. Hence,

$$\begin{aligned}
\dim(E(\mathcal{U}+\mathcal{W}) + M(\mathcal{U}+\mathcal{W})) &= \dim(E(\mathcal{U}^*+\mathcal{W}^*) + M(\mathcal{U}^*+\mathcal{W}^*)) \\
&\geqq \dim(E\mathcal{W}^* + M\mathcal{U}^*) \\
&= \dim E\mathcal{W}^* + \dim M\mathcal{U}^* - \dim(E\mathcal{W}^* \cap M\mathcal{U}^*) \\
&= \dim \mathcal{W}^* + \dim \mathcal{U}^* - \dim \mathcal{U}^* \cap \mathcal{W}^* \\
&= \dim(\mathcal{U}^*+\mathcal{W}^*) \\
&= \dim(\mathcal{U}+\mathcal{W}).
\end{aligned}$$

It now follows from Lemma 3.1(i) and Theorem 3.2 of [8] that the subspace denoted by H_N in [8] will serve as $\mathcal{Q}$.

This completes the proof of Lemma 3.3.

Proof of Theorem 3.2. It is easy to see that

$$E\mathcal{W} + M\mathcal{U} = E(\mathcal{W}+\mathcal{U}) + M(\mathcal{W}+\mathcal{U}) = E\mathcal{W}^* + M\mathcal{Q}$$

and so

$$f(t) = Eg(t) + Mh(t) \quad \text{for all } t \geqq 0$$

where $g(t) \in \mathcal{W}^*$ and $h(t) \in \mathcal{Q}$ for all $t \geqq 0$. Now we split (6) into two parts:

$$E\dot{v}(t) = Mv(t) + Eg(t), \qquad v(0) = x_0 - q(0), \tag{19}$$

$$E\dot{q}(t) = Mq(t) + Mh(t) \tag{20}$$

where $v(t) \in \mathcal{W}^*$ and $q(t) \in \mathcal{Q}$ for all $t \geqq 0$. Clearly, if we can solve (19) and (20), then the sum of the solutions satisfies (6). To solve (20) let $K : \mathcal{X} \to \mathcal{X}$ be (the matrix of) a linear transformation such that

$$K\mathcal{Q} \subseteq \mathcal{Q}, \quad MKu = Eu \quad \text{for all } u \in \mathcal{Q}.$$

It is shown in Lemma 3.1(iii) of [8] that K is nilpotent on $\mathcal{Q}$, say $K^s = 0$. Now it is easy to see, by direct substitution, that a solution of (20) is

$$q(t) = -\sum_{i=0}^{s-1} \frac{d^i}{dt^i} K^i h(t).$$

Note that the solution determines $q(0)$, so the omission of initial conditions from (20) was deliberate.

Equation (19) has a solution if and only if $v(0) \in \mathcal{W}^*$, so this determines the affine subspace of $\mathcal{X}$ referred to in the statement of the theorem. Then it is easy to see, again by direct substitution, that

$$v(t) = e^{Lt}v_0 + \int_0^t e^{L(t-\tau)}g(\tau)\,d\tau,$$

where L is the matrix introduced in the proof of Lemma 3.1, satisfies (19).

This completes the proof of Theorem 3.2.

Now we can easily establish that the existence of subspaces $\mathcal{U}$, $\mathcal{V}$, $\mathcal{Z}$ as specified in the Main Theorem is sufficient for the existence of F achieving disturbance decoupling. Indeed, if F satisfies the requirements of Theorem 2.1 and Lemma 2.2 with respect to these subspaces, then the hypotheses of Theorem 3.2 are satisfied with $M = A - BF$ and $f(t) = Sd(t)$. Thus, (15) has a solution contained in $\mathcal{U} + \mathcal{W} = \mathcal{U} + \mathcal{V} \subseteq \ker C$ whatever the function $d(t)$ provided that it is sufficiently differentiable and that x_0 is appropriately related to the value of d and its derivatives at time $t = 0$.

This completes the proof of sufficiency in the Main Theorem.

4. Proof of the Main Theorem: necessity. In this section we complete the proof of the Main Theorem by outlining a proof that the conditions for disturbance decoupling are necessary, as well as sufficient. Suppose F is a feedback achieving disturbance decoupling in the closed loop system

$$\begin{aligned} &E\dot{x} = (A - BF)x + Sd(t), \qquad x(0) = x_0, \\ &y = Cx. \end{aligned} \tag{21}$$

Let $\mathcal{X}_0 \subseteq \ker C$ be the subspace of $\mathcal{X}$ minimal subject to containing every solution of (21) for which $y(t) \equiv 0$. It is clear that if $x(t) \in \mathcal{X}_0$ for all $t \geqq 0$, then $\dot{x}(t) \in \mathcal{X}_0$ for all $t \geqq 0$. Since F achieves disturbance decoupling

$$\operatorname{Im} S \subseteq E\mathcal{X}_0 + (A - BF)\mathcal{X}_0$$

and we have Lemma 4.1.

LEMMA 4.1. *To prove the necessity of the conditions in the Main Theorem it is enough to show that*

$$\mathcal{X}_0 = \mathcal{U} \oplus \mathcal{V}$$

for some subspaces $\mathcal{U}$, $\mathcal{V}$ satisfying

$$E\mathcal{U} \subseteq (A - BF)\mathcal{U}, \qquad (A - BF)\mathcal{V} \subseteq E\mathcal{V}.$$

Proof. If the conditions of this lemma hold, then

$$\begin{aligned} \operatorname{Im} S &\subseteq E(\mathcal{U} + \mathcal{V}) + (A - BF)(\mathcal{U} + \mathcal{V}) \\ &\subseteq E\mathcal{V} + (A - BF)\mathcal{U}. \end{aligned}$$

It is easy to see that $\mathcal{Z} = (A - BF)\mathcal{U}$ satisfies (a) and (b) of the Main Theorem.

This completes the proof of Lemma 4.1.

We sketch a construction of the subspaces $\mathcal{U}$, $\mathcal{V}$ referred to here using the Kronecker canonical form $A_1 - \lambda E_1$ of the pencil $A_0 - \lambda E_0$, where A_0 and E_0 are the matrices of the linear mappings obtained by restricting $A - BF$ and E, respectively, to $\mathcal{X}_0$. It will be convenient to refer to [7] for details of the Kronecker canonical form as Wilkinson is directly concerned with the differential equations

$$E_1\dot{x}(t) = A_1x(t) + f(t).$$

We recall that we may be dealing with a singular pencil since A_0 and E_0 might not be square matrices. We begin therefore by referring to [7, p. 249], where

> ... the essential features introduced by singularity (are) summarised as follows:
> Corresponding to each (of the minimal column indices) the general solution contains one arbitrary function ...
> Corresponding to each (of the minimal row indices) $\eta \cdots$ a compatibility relation is required between
>
> $$f_i, Df_{i+1}, D^2 f_{i+2}, \cdots, D^\eta f_{i+\eta} \left(D = \frac{d}{dt} \right)$$
>
> ...

Since we are insisting on uniqueness of solutions, $A_1 - \lambda E_1$ contains no minimal column indices and, as we are allowing any sufficiently differentiable $d(t)$ without any compatibility requirement, all the minimal row indices are zero. Thus there exist nonsingular matrices P and Q such that

$$P(A_0 - \lambda E_0)Q = A_1 - \lambda E_1 = \begin{pmatrix} 0 \\ A_2 - \lambda E_2 \end{pmatrix}$$

where $A_2 - \lambda E_2$ is a regular pencil. Furthermore, $A_2 - \lambda E_2$ can be taken to be block diagonal with diagonal blocks given either by

$$I - \lambda J(0) \tag{22}$$

or by

$$J(\lambda_i) - \lambda I \tag{23}$$

where "$J(\cdots)$" denotes "a Jordan matrix with eigenvalue ..." and λ_i ranges over the finite eigenvalues of the pencil $A_0 - \lambda E_0$. We will assume that the diagonal blocks are arranged so that all those of the first kind appear above and to the left of any of the second kind. Thus

$$P(A_0 - \lambda E_0)Q = \begin{pmatrix} 0 & 0 \\ A_3 - \lambda E_3 & 0 \\ 0 & A_4 - \lambda E_4 \end{pmatrix} \tag{24}$$

where all the infinite eigenvalues of $A_0 - \lambda E_0$ are accounted for in $A_3 - \lambda E_3$ and the finite eigenvalues in $A_4 - \lambda E_4$.

The matrices A_1 and E_1 may be taken to act on the space of column vectors given by $\mathscr{X}_1 = Q^{-1}\mathscr{X}_0$. Let $\mathscr{U}_1$, $\mathscr{V}_1$ denote subspaces of $\mathscr{X}_1$ corresponding to a partition of these column vectors conformable to that in (24). Then it is clear that

$$\mathscr{X}_1 = \mathscr{U}_1 \oplus \mathscr{V}_1, \quad E_1\mathscr{U}_1 \subseteq A_1\mathscr{U}_1, \quad A_1\mathscr{V}_1 \subseteq E_1\mathscr{V}_1$$

and so $\mathscr{U} = Q\mathscr{U}_1$ and $\mathscr{V} = Q\mathscr{V}_1$ satisfy the requirements of Lemma 4.1.

This completes the proof of the necessity of the conditions in the Main Theorem, so we have now proved the Main Theorem in its entirety.

5. Conclusion. We have provided necessary and sufficient conditions for the existence of a state feedback that achieves disturbance decoupling in a linear time-invariant descriptor system. We have taken care to ensure that the resulting closed loop system has smooth solutions for a wide class of disturbance functions and initial conditions and that, when a solution exists, it is unique. Although different application will require different stipulations about the existence of solutions, "control" in the context of descriptor systems seems almost certain to require uniqueness of solutions.

Our mode of argument has been in the spirit of Wonham [9] and the structure of our proof clearly resembles that of the corresponding result in the state space case. However, we have found the usual presentation, in terms of the properties of uniquely determined maximal elements of certain collections of subspaces, neither possible nor essential in the descriptor case. This has meant that our proofs are less constructive than we would like; some work is needed to integrate our notions with other, more algorithmic, approaches. It would be particularly interesting to identify, in a constructive way, the maximal subspace that is the image under E of a subspace $\mathcal{V}$ of $\ker C$ satisfying (i) and (ii) of Definition 1.3. Lemma 2.4 shows that this subspace is uniquely determined.

Throughout this paper we have had in mind a linear system described by a mixture of algebraic and *differential* equations; the extension of our results to the discrete-time case should not present any major difficulty.

Acknowledgment. It is a pleasure to acknowledge that detailed comments from referees of an earlier version of this paper led to clearer thinking and sharper results.

REFERENCES

[1] L. R. FLETCHER, *Regularisability of descriptor systems*, Internat. J. Systems Sci., 17 (1986), pp. 843–847.

[2] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Invariant Subspaces of Matrices with Applications*, Canadian Mathematical Society Series of Monographs and Advanced Texts, John Wiley, New York, 1986.

[3] F. L. LEWIS, *A survey of linear singular systems*, J. Circuits Systems Signal Process., 5 (1986), pp. 3–36.

[4] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312–321.

[5] A. MANITIUS, *F-controllability and observability of linear retarded systems*, Appl. Math. Optim., 19 (1982), pp. 73–95.

[6] G. C. VERGHESE, B. C. LEVY, AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 811–831.

[7] J. H. WILKINSON, *Linear differential equations and Kronecker's canonical form*, in Proc. Symposium on Recent Advances in Numerical Analysis, C. de Boor and G. H. Golub, eds., Academic Press, New York, 1978.

[8] K. T. WONG, *The eigenvalue problem $\lambda Tx + Sx$*, J. Differential Equations, 16 (1974), pp. 270–280.

[9] W. M. WONHAM, *Linear Multivariable Control—a Geometric Approach*, Second edition, Springer-Verlag, New York, 1979.

[10] Z. ZHENG, 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. 27, No. 6, pp. 1333-1360, November 1989

006

EXACT PENALTY FUNCTIONS IN CONSTRAINED OPTIMIZATION*

G. DI PILLO† AND L. GRIPPO‡

Abstract. In this paper formal definitions of exactness for penalty functions are introduced and sufficient conditions for a penalty function to be exact according to these definitions are stated, thus providing a unified framework for the study of both nondifferentiable and continuously differentiable penalty functions. In this framework the best-known classes of exact penalty functions are analyzed, and new results are established concerning the correspondence between the solutions of the constrained problem and the unconstrained minimizers of the penalty functions.

Key words. exact penalty functions, nonlinear programming, constrained optimization

AMS(MOS) subject classifications. 49D30, 49D37, 90C30

1. Introduction.

A considerable amount of investigation, both from the theoretical and the computational point of view, has been devoted to methods that attempt to solve nonlinear programming problems by means of a single minimization of an unconstrained function. Methods of this kind are usually termed *exact penalty methods*, as opposed to the *sequential penalty methods*, which include the quadratic penalty method and the method of multipliers (see, e.g., [4], [23], and [26]).

We can subdivide exact penalty methods into two classes: methods based on *exact penalty functions* and methods based on *exact augmented Lagrangian functions.* In our terminology, the term "exact penalty function" is used when the variables of the unconstrained problem are in the same space as the variables of the original constrained problem, whereas the term "exact augmented Lagrangian function" is used when the unconstrained problem has to be minimized on the product space of the problem variables and of the multipliers.

Exact penalty functions can be subdivided, in turn, into two main classes: *nondifferentiable* exact penalty functions and *continuously differentiable* exact penalty functions.

Nondifferentiable exact penalty functions were introduced for the first time in [39] and have been widely investigated in recent years (see, e.g., [1], [2], [5]-[10], [22], [25], [29], and [35]). Continuously differentiable exact penalty functions were introduced in [24] for equality constrained problems and in [28] for problems with inequality constraints; further contributions have been given in [14], [15], and [34].

Exact augmented Lagrangian functions were introduced in [11] and [12] and have been further investigated in [3], [4], [19]-[21], [31], and [38].

In this paper we restrict our attention to exact penalty functions, with the aim of providing a unified framework which applies both to the nondifferentiable and to the continuously differentiable case.

We start from the introduction of formal definitions of various kinds of exactness that attempt to capture the most relevant aspects of the notion of exactness in the context of constrained optimization. This is motivated by the fact that in the current

* Received by the editors June 6, 1988; accepted for publication (in revised form) November 11, 1988. This research was partially supported by the National Research Program on "Modelli e Algoritmi per l'Ottimizzazione," Ministero della Pubblica Istruzione, Rome, Italy.

† Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza," Via Eudossiana 18, 00184 Rome, Italy.

‡ Istituto di Analisi dei Sistemi ed Informatica del Consiglio Nazionale delle Ricerche, Viale Manzoni 30, 00185 Rome, Italy.

literature the term *exact penalty function* seems to be used without a definite agreement on its meaning. In particular, as noted in [29], most of the literature on this subject is mainly concerned with conditions that ensure that the penalty function has a local (global) minimum at a local (global) minimum point of the constrained problem. On the other hand, since the penalty approach is an attempt to solve a constrained problem by the minimization of an unconstrained function, this characterization is fully satisfactory only when both the constrained problem and the penalty function are convex. In the nonconvex case, the study of converse properties appears to be of greater interest, as they ensure that local (global) minimizers of the penalty function are local (global) solutions of the constrained problem.

Moreover, again in the nonconvex case, a distinction has to be made between properties of exactness pertaining to global solutions and properties pertaining to local solutions. It will be shown that, for the same penalty function, different kinds of exactness can be established under different regularity requirements on the problem constraints.

Finally, the correspondence between the constrained and the unconstrained minimization problem can only be established with reference to a compact set containing the problem solutions, and this must be carefully taken into account in the analysis of the properties of exactness.

The formal definitions mentioned so far constitute the basis for the development of sufficient conditions for a penalty function to be exact according to some specified notion of exactness. In particular, we establish sufficient conditions which apply both to the nondifferentiable and to the continuously differentiable case, thus providing a unified framework for the analysis and the construction of exact penalty functions. In this framework, we consider the best-known classes of exact penalty functions, and we provide a complete analysis of their properties, recovering known results and establishing new ones.

The paper is organized as follows. Section 2 contains the problem statement, basic notation, and preliminary results. In § 3 we formalize the definitions of various kinds of exactness of penalty functions, which are classified as *weak exactness*, *exactness*, *strong exactness*, and *global* (*weak*, *strong*) *exactness*. Section 4 deals with nondifferentiable penalty functions: we analyze the properties of l_q exact penalty functions as well as those of the globally exact nondifferentiable penalty function considered in [16]. In § 5 we study continuously differentiable exact penalty functions, and we introduce a globally exact differentiable penalty function for mixed equality and inequality constrained problems by extending the results given in [15].

Computational aspects are beyond the scope of this paper. We refer, e.g., to [3], [4], [9], [18], [21], [27], [28], [33], [34], [36], and [37] for some algorithmic applications of exact penalty functions.

2. Problem statement, basic notation and preliminary results. The problem considered here is the general nonlinear programming problem:

$$\text{(P)} \qquad \begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g(x) \leqq 0, \qquad h(x) = 0, \end{aligned}$$

where $f: \mathbb{R}^n \to \mathbb{R}$, $g: \mathbb{R}^n \to \mathbb{R}^m$, $h: \mathbb{R}^n \to \mathbb{R}^p$, $p \leqq n$ are continuously differentiable functions and the feasible set

$$\mathcal{F} := \{x \in \mathbb{R}^n : g(x) \leqq 0, h(x) = 0\}$$

is assumed to be nonempty.

We denote by $\mathcal{G}_{\mathcal{P}}$ and $\mathcal{L}_{\mathcal{P}}$, respectively, the set of global solutions and the set of local solutions of problem (P) and we assume that $\mathcal{G}_{\mathcal{P}}$ is nonempty.

The Lagrangian function associated with problem (P) is the function $L: \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ defined by

$$L(x, \lambda, \mu) := f(x) + \lambda' g(x) + \mu' h(x).$$

A *Kuhn-Tucker triple* for problem (P) is a triple $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^p$ such that

$$\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0,$$
$$\bar{\lambda}' g(\bar{x}) = 0,$$
$$\bar{\lambda} \geqq 0,$$
$$g(\bar{x}) \leqq 0,$$
$$h(\bar{x}) = 0.$$

We denote by $\mathcal{T}$ the set

$$\mathcal{T} := \{\bar{x} \in \mathbb{R}^n : \text{there exist } (\bar{\lambda}, \bar{\mu}) \text{ such that}$$
$$(\bar{x}, \bar{\lambda}, \bar{\mu}) \text{ is a K-T triple for problem (P)}\}.$$

For any $x \in \mathbb{R}^n$ we define the index sets:

$$I_0(x) := \{i : g_i(x) = 0\}$$
$$I_+(x) := (i : g_i(x) \geqq 0\}.$$

We adopt the following terminology.

The *linear independence constraint qualification* (LICQ) holds at $x \in \mathbb{R}^n$ if the gradients $\nabla g_i(x)$, $i \in I_0(x)$, $\nabla h_j(x)$, $j = 1, \cdots, p$ are linearly independent.

The *Mangasarian-Fromovitz constraint qualification* (MFCQ) holds at $x \in \mathbb{R}^n$ if $\nabla h_j(x)$, $j = 1, \cdots, p$ are linearly independent and there exists a $z \in \mathbb{R}^n$ such that

$$\nabla g_i(x)' z < 0, \qquad i \in I_0(x)$$
$$\nabla h_j(x)' z = 0, \qquad j = 1, \cdots, p.$$

It can be shown, by using the theorems of the alternative [32] that the MFCQ can be restated as follows.

The MFCQ holds at $x \in \mathbb{R}^n$ if there exist no u_i, $i \in I_0(x)$, and v_j, $j = 1, \cdots, p$ such that

$$\sum_{i \in I_0(x)} u_i \nabla g_i(x) + \sum_{j=1}^{p} v_j \nabla h_j(x) = 0,$$
$$u_i \geqq 0, \qquad i \in I_0(x),$$
$$(u_i, i \in I_0(x), v_j, \qquad j = 1, \cdots, p) \neq 0.$$

In some cases we shall make use of a stronger constraint qualification, which is stated in the following equivalent formulations.

The *extended Mangasarian-Fromovitz constraint qualification* (EMFCQ) holds at $x \in \mathbb{R}^n$ if $\nabla h_j(x)$, $j = 1, \cdots, p$ are linearly independent and there exists a $z \in \mathbb{R}^n$ such that

$$\nabla g_i(x)' z < 0, \qquad i \in I_+(x)$$
$$\nabla h_j(x)' z = 0, \qquad j = 1, \cdots, p.$$

The EMFCQ holds at $x \in \mathbb{R}^n$ if there exist no u_i, $i \in I_+(x)$, and v_j, $j = 1, \cdots, p$ such that

$$\sum_{i \in I_+(x)} u_i \nabla g_i(x) + \sum_{j=1}^{p} v_j \nabla h_j(x) = 0,$$

$$u_i \geqq 0, \qquad i \in I_+(x),$$

$$(u_i, i \in I_+(x), v_j, \qquad j = 1, \cdots, p) \neq 0.$$

It can be noted that the LICQ implies the MFCQ and that the EMFCQ implies the MFCQ.

It is known that if $\bar{x}$ is a local solution of problem (P) and if the MFCQ holds at $\bar{x}$, then $\bar{x} \in \mathcal{T}$, that is, there exist K-T multipliers $(\bar{\lambda}, \bar{\mu})$ associated with $\bar{x}$.

We recall that a nonempty set $\mathcal{C}^* \subseteq \mathcal{C}$ is called an *isolated set* of $\mathcal{C}$ if there exists a closed set $\mathcal{H}$ such that $\mathcal{C}^*$ is contained in the interior $\mathring{\mathcal{H}}$ of $\mathcal{H}$ and such that if $x \in \mathcal{H} - \mathcal{C}^*$, then $x \notin \mathcal{C}$. Isolated sets of local minimum points possess the property stated in the following lemma, which is proved in [23].

LEMMA 1. *Let $\mathcal{C}^*$ be an isolated compact set of local minimum points of problem* (P), *corresponding to the local minimum value f^*; then there exists a compact set $\mathcal{H} \subset \mathbb{R}^n$, such that $\mathcal{C}^* \subset \mathring{\mathcal{H}}$, and for any point $x \in \mathcal{H} \cap \mathcal{F}$, if $x \notin \mathcal{C}^*$, then $f(x) > f^*$.*

We also state the following lemma, which for $q \geqq 2$ is an obvious consequence of the equivalence of the norms $\| \cdot \|_q$ and $\| \cdot \|_{q-1}$ on $\mathbb{R}^n$.

LEMMA 2. *Let $q \in \mathbb{R}$, $1 < q < \infty$. Then, there exists a number $\mu > 0$ such that for all $z \in \mathbb{R}^n$, we have*:

$$\sum_{i=1}^{n} |z_i|^{q-1} \geqq \mu \|z\|_q^{q-1}.$$

Proof. The assertion follows from a more general result on positive homogeneous continuous functions ([30, Thm. 5.4.4]). □

In the sequel we shall be concerned with compact perturbations of the feasible set. In particular, we shall consider the case in which $\mathcal{F}$ is compact and there exists a vector $\beta = (\alpha_0, \alpha')'$ with $\alpha_0 \in \mathbb{R}$, $\alpha \in \mathbb{R}^m$, $\beta > 0$, such that the set

$$\mathcal{S}_\beta := \{x \in \mathbb{R}^n : g(x) \leqq \alpha, \|h(x)\|_2^2 \leqq \alpha_0\}$$

obtained by relaxing the problem constraints is compact. It can be shown, by extending a similar result given in [15] for the inequality constrained case, that under the following assumptions:

(i) there exists a $\bar{\beta} \in \mathbb{R}^{m+1}$, $\bar{\beta} > 0$, such that $\mathcal{S}_{\bar{\beta}}$ is compact,

(ii) the MFCQ holds on $\mathcal{F}$,

there exists a compact set S_β, with $\beta > 0$, where the EMFCQ is satisfied.

We make use of the following notation. Given the set $\mathcal{A}$, we denote by $\mathring{\mathcal{A}}$, $\partial\mathcal{A}$, and $\bar{\mathcal{A}}$, respectively, the interior, the boundary, and the closure of $\mathcal{A}$. Given a vector u with components u_i, $i = 1, \cdots, m$ we denote by u^+ the vector with components:

$$u_i^+ := \max [0, u_i], \qquad i = 1, \cdots, m$$

and by U the diagonal matrix defined by:

$$U := \operatorname{diag}(u_i), \qquad i = 1, \cdots, m.$$

Given a function $F : \mathbb{R}^n \to \mathbb{R}$, we denote by $DF(x, d)$ the directional derivative of F at x along the direction d. We say that $\bar{x}$ is a *critical point* of F if $DF(x, d) \geqq 0$ for all $d \in \mathbb{R}^n$. If $\bar{x}$ is a critical point of F and F is differentiable at $\bar{x}$, we have $\nabla F(\bar{x}) = 0$; in this case we say that $\bar{x}$ is a *stationary point* of F.

Finally, we denote by $\mathcal{B}(\bar{x}; \rho)$ the open ball around $\bar{x}$ with radius $\rho > 0$.

3. Definitions of exactness for penalty functions. Roughly speaking, an *exact penalty function* for problem (P) is a function $F(x;\varepsilon)$, where $\varepsilon>0$ is a *penalty parameter*, with the property that there is an appropriate parameter choice such that a single unconstrained minimization of $F(x;\varepsilon)$ yields a solution to problem (P). In particular, we require that there is an *easy* way for finding correct parameter values by imposing that exactness is retained for all ε ranging on some set of nonzero measure. More specifically, we take $\varepsilon\in(0,\varepsilon^*]$ where $\varepsilon^*>0$ is a suitable *threshold value.*

In practice, the existence of a threshold value for the parameter ε, and hence the possibility of constructing the exact penalty function $F(x;\varepsilon)$, can only be established with reference to some compact set $\mathcal{D}$. Therefore, instead of problem (P) we shall consider the following problem.

$$(\tilde{\mathrm{P}}) \qquad \text{minimize } f(x), \qquad x\in\mathcal{F}\cap\mathcal{D},$$

where $\mathcal{D}$ is a compact subset of $\mathbb{R}^n$ such that $\mathcal{F}\cap\mathring{\mathcal{D}}\neq\varnothing$. It can be observed that if $\mathcal{F}\subset\mathcal{D}$, then problem ($\tilde{\mathrm{P}}$) and problem (P) are equivalent.

We denote by $\mathcal{G}_{\tilde{\mathcal{P}}}$ and $\mathcal{L}_{\tilde{\mathcal{P}}}$, respectively, the set of global solutions and the set of local solutions of problem ($\tilde{\mathrm{P}}$), that is:

$$\mathcal{G}_{\tilde{\mathcal{P}}}:=\{x\in\mathcal{F}\cap\mathcal{D}: f(x)\leqq f(y), \text{ for all } y\in\mathcal{F}\cap\mathcal{D}\}$$

$$\mathcal{L}_{\tilde{\mathcal{P}}}:=\{x\in\mathcal{F}\cap\mathcal{D}: \text{ for some } \rho>0\ f(x)\leqq f(y), \text{ for all } y\in\mathcal{F}\cap\mathcal{D}\cap\mathcal{B}(x;\rho)\}.$$

We have, obviously, that $\mathcal{L}_{\tilde{\mathcal{P}}}\cap\mathring{\mathcal{D}}=\mathcal{L}_{\mathcal{P}}\cap\mathring{\mathcal{D}}$; moreover, if $\mathcal{G}_{\mathcal{P}}\cap\mathring{\mathcal{D}}\neq\varnothing$, we have also $\mathcal{G}_{\mathcal{P}}\cap\mathring{\mathcal{D}}=\mathcal{G}_{\tilde{\mathcal{P}}}\cap\mathring{\mathcal{D}}$.

For any given $\varepsilon>0$, let $F(x;\varepsilon)$ be a continuous real function defined on a set $\mathcal{E}$, such that $\mathring{\mathcal{D}}\subseteq\mathcal{E}\subseteq\mathcal{D}$ and consider the following problem.

$$(\mathrm{Q}) \qquad \text{minimize } F(x;\varepsilon), \qquad x\in\mathring{\mathcal{D}}.$$

Since $\mathring{\mathcal{D}}$ is an open set, any local solution of problem (Q), provided it exists, is unconstrained; thus problem (Q) can be considered as an essentially unconstrained problem. The sets of global and local solutions of problem (Q) are denoted, respectively, by $\mathcal{G}_{\mathcal{Q}}(\varepsilon)$ and $\mathcal{L}_{\mathcal{Q}}(\varepsilon)$:

$$\mathcal{G}_{\mathcal{Q}}(\varepsilon):=\{x\in\mathring{\mathcal{D}}: F(x;\varepsilon)\leqq F(y;\varepsilon), \text{for all } y\in\mathring{\mathcal{D}}\}$$

$$\mathcal{L}_{\mathcal{Q}}(\varepsilon):=\{x\in\mathring{\mathcal{D}}: \text{ for some } \rho>0\ F(x;\varepsilon)\leqq F(y;\varepsilon), \text{for all } y\in\mathring{\mathcal{D}}\cap\mathcal{B}(x;\rho)\}.$$

There are different kinds of relationships between problem ($\tilde{\mathrm{P}}$) and problem (Q), which can be associated with different notions of exactness.

A first possibility is that of considering a correspondence between global minimizers of problem ($\tilde{\mathrm{P}}$) and global minimizers of problem (Q). This correspondence is established formally in the following definition.

DEFINITION 1. We say that the function $F(x;\varepsilon)$ is a *weakly exact penalty function* for problem (P) with respect to the set $\mathcal{D}$ if there exists an $\varepsilon^*>0$ such that, for all $\varepsilon\in(0,\varepsilon^*]$, any global solution of problem ($\tilde{\mathrm{P}}$) is a global minimum point of problem (Q) and conversely; that is if for some $\varepsilon^*>0$:

$$\mathcal{G}_{\tilde{\mathcal{P}}}=\mathcal{G}_{\mathcal{Q}}(\varepsilon), \quad \text{for all } \varepsilon\in(0,\varepsilon^*].$$

The property stated above guarantees that the constrained problem can actually be solved over $\mathcal{D}$ by means of the *global unconstrained minimization* of $F(x;\varepsilon)$ for sufficiently small values of the parameter ε.

We remark that if all global solutions of problem (P) are contained in $\mathring{\mathcal{D}}$, then problem (P) and problem ($\tilde{\mathrm{P}}$) possess the same global solutions. In this case, weak exactness implies that global solutions of problem (P) and global minimizers of problem (Q) are the same.

The notion of exactness expressed by Definition 1 appears to be of limited value for general nonlinear programming problems, since it does not give a meaning to local minimizers of the penalty function, while unconstrained minimization algorithms determine only local minimizers. Therefore, we introduce a further requirement concerning local minimizers which gives rise to a stronger notion of exactness.

DEFINITION 2. We say that the function $F(x; \varepsilon)$ is an *exact penalty function* for problem (P) with respect to the set $\mathcal{D}$ if there exists an $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*]$, $\mathcal{G}_{\mathcal{P}} = \mathcal{G}_{\mathcal{Q}}(\varepsilon)$ and, moreover, any local unconstrained minimizer of problem (Q) is a local solution of problem (P), that is:

$$\mathcal{L}_{\mathcal{Q}}(\varepsilon) \subseteq \mathcal{L}_{\mathcal{P}}, \quad \text{for all } \varepsilon \in (0, \varepsilon^*].$$

It must be remarked that the notion of exactness given in Definition 2 does not require that all local solutions of problem (P) in $\mathring{\mathcal{D}}$ correspond to local minimizers of the exact penalty functions. A one-to-one correspondence of local minimizers does not seem to be required, in practice, to give a meaning to the notion of exactness, since the condition $\mathcal{G}_{\mathcal{P}} = \mathcal{G}_{\mathcal{Q}}(\varepsilon)$ ensures that global solutions of problem ($\tilde{\mathrm{P}}$) are preserved. However, for the classes of exact penalty functions considered in the sequel, it will be shown that this correspondence can be established, also, at least with reference to isolated compact sets of local minimizers of problem (P) contained in $\mathring{\mathcal{D}}$. Thus, we can also consider the following definition.

DEFINITION 3. We say that the function $F(x; \varepsilon)$ is a *strongly exact penalty function* for problem (P) with respect to the set $\mathcal{D}$ if there exists an $\varepsilon^* > 0$ such that, for all $\varepsilon \in (0, \varepsilon^*]$, $\mathcal{G}_{\mathcal{P}} = \mathcal{G}_{\mathcal{Q}}(\varepsilon)$, $\mathcal{L}_{\mathcal{Q}}(\varepsilon) \subseteq \mathcal{L}_{\mathcal{P}}$, and, moreover, any local solution of problem (P) belonging to $\mathring{\mathcal{D}}$ is a local unconstrained minimizer of $F(x; \varepsilon)$, that is:

$$\mathcal{L}_{\mathcal{P}} \cap \mathring{\mathcal{D}} \subseteq \mathcal{L}_{\mathcal{Q}}(\varepsilon) \quad \text{for all } \varepsilon \in (0, \varepsilon^*].$$

The properties considered in the preceding definitions do not characterize the behavior of $F(x; \varepsilon)$ on the boundary of $\mathring{\mathcal{D}}$. Although this may be irrelevant from the conceptual point of view in connection with the notion of exactness, it may assume a considerable interest from the computational point of view, when unconstrained descent methods are employed for the minimization of $F(x; \varepsilon)$. In fact, it may happen that there exist points of $\mathring{\mathcal{D}}$ such that a descent path for $F(x; \varepsilon)$ that originates at some of these points crosses the boundary of $\mathring{\mathcal{D}}$. This implies that the sequence of points produced by an unconstrained algorithm may be attracted toward a stationary point of $F(x; \varepsilon)$ out of $\mathring{\mathcal{D}}$ or may not admit a limit point. Therefore, it could be difficult to construct minimizing sequences for $F(x; \varepsilon)$ which are globally convergent on $\mathring{\mathcal{D}}$ toward the solutions of the constrained problem. In order to avoid this difficulty, it is necessary to impose further conditions on $F(x; \varepsilon)$, and we are led to introduce the notion of *global exactness* of a penalty function.

DEFINITION 4. The function $F(x; \varepsilon)$ is said to be a *globally* (*weakly, strongly*) *exact penalty function* for problem (P) with respect to the set $\mathcal{D}$ if it is (weakly, strongly) exact and, moreover, for any $\varepsilon > 0$ and for any $\hat{x} \in \partial\mathcal{D}$ there exists a neighborhood $\mathcal{B}(\hat{x}, \rho)$ such that if $\{x_k\} \subset \mathring{\mathcal{D}}$ and $\lim_{k\to\infty} x_k = \hat{x}$, we have:

$$\liminf_{k\to\infty} F(x_k; \varepsilon) > F(x; \varepsilon),$$

for all $x \in \mathcal{B}(\hat{x}; \rho) \cap \mathring{\mathcal{D}}$.

The condition given above excludes the existence of minimizing sequences for $F(x; \varepsilon)$ originating in $\mathring{\mathcal{D}}$ that have limit points on the boundary. In fact, we can state the following proposition.

PROPOSITION 1. *Let $F(x;\varepsilon)$ be a (weakly, strongly) globally exact penalty function with respect to the set $\mathscr{D}$ and let $\{x_k\}\subset \mathring{\mathscr{D}}$ be a sequence such that $F(x_{k+1};\varepsilon)\leqq F(x_k;\varepsilon)$. Then, any limit point of $\{x_k\}$ belongs to $\mathring{\mathscr{D}}$.*

Proof. By the compactness of $\mathscr{D}$ there exists a subsequence, which we relabel $\{x_k\}$, such that $x_k\to\hat{x}\in\mathscr{D}$. Reasoning by contradiction, assume that $\hat{x}\in\partial\mathscr{D}$. Then, recalling Definition 4, we have, for sufficiently large values of j, $\liminf_{k\to\infty} F(x_k;\varepsilon)>F(x_j;\varepsilon)$, which contradicts the assumption $F(x_k;\varepsilon)\leqq F(x_j;\varepsilon)$ for all $k\geqq j$. □

4. Sufficient conditions for exactness. In this section we state sufficient conditions which imply that a penalty function $F(x;\varepsilon)$ possesses some of the properties of exactness considered in the preceding section. Everywhere below we suppose that the following assumption holds.

Assumption (A1). Any global solution of problem $(\tilde{\mathrm{P}})$ belongs to the set $\mathring{\mathscr{D}}$, that is: $\mathscr{G}_{\tilde{\mathscr{P}}}\subset\mathring{\mathscr{D}}$.

We note that Assumption (A1) concerns the selection of the set $\mathscr{D}$ and implies that $\mathscr{G}_{\tilde{\mathscr{P}}}\subset\mathscr{L}_{\mathscr{P}}$; it can be satisfied, in particular, by a proper choice of $\mathscr{D}$, whenever the global solutions of problem (P) belong to a bounded subset of $\mathscr{F}$.

Let $\mathscr{K}$ be the subset of $\mathscr{F}\cap\mathring{\mathscr{D}}$ where the function $F(x;\varepsilon)$ takes the same values of $f(x)$, that is:

$$\mathscr{K}:=\{x\in\mathscr{F}\cap\mathring{\mathscr{D}}\colon F(x;\varepsilon)=f(x)\text{ for all }\varepsilon>0\}. \tag{1}$$

The next theorem establishes a sufficient condition for $F(x;\varepsilon)$ to be a weakly exact penalty function in the sense of Definition 1.

THEOREM 1. *Let $F(x;\varepsilon)$ be such that the following conditions are satisfied.*

(a_1) *For any $\varepsilon>0$, the function $F(x;\varepsilon)$ admits a global minimum point on a set $\mathscr{E}$, such that $\mathring{\mathscr{D}}\subseteq\mathscr{E}\subseteq\mathscr{D}$.*

(a_2) *If $\{\varepsilon_k\}$ and $\{x_k\}\subseteq\mathscr{E}$ are sequences such that $\lim_{k\to\infty}\varepsilon_k=0$, $\lim_{k\to\infty}x_k=\hat{x}\in\mathscr{D}$ and $\limsup_{k\to\infty}F(x_k,\varepsilon_k)<\infty$, we have $\hat{x}\in\mathscr{F}\cap\mathscr{D}$ and $f(\hat{x})\leqq\limsup_{k\to\infty}F(x_k,\varepsilon_k)$.*

(a_3) $\mathscr{G}_{\tilde{\mathscr{P}}}\subseteq\mathscr{K}$.

(a_4) *For any $\hat{x}\in\mathscr{G}_{\tilde{\mathscr{P}}}$ there exist numbers $\varepsilon(\hat{x})>0$ and $\sigma(\hat{x})>0$ such that, for all $\varepsilon\in(0,\varepsilon(\hat{x})]$, if $\mathscr{G}_{\mathscr{Q}}(\varepsilon)\neq\varnothing$ and $x_\varepsilon\in\mathscr{G}_{\mathscr{Q}}(\varepsilon)$ is a global minimum point of problem* (Q) *satisfying $\|x_\varepsilon-\hat{x}\|\leqq\sigma(\hat{x})$, we have $x_\varepsilon\in\mathscr{K}$.*

Then, $F(x;\varepsilon)$ is a weakly exact penalty function for problem (P) *with respect to the set $\mathscr{D}$.*

Proof. We show first that there exists an $\varepsilon^*>0$ such that, for all $\varepsilon\in(0,\varepsilon^*]$ we have $\mathscr{G}_{\mathscr{Q}}(\varepsilon)\neq\varnothing$ and $\mathscr{G}_{\mathscr{Q}}(\varepsilon)\subseteq\mathscr{G}_{\tilde{\mathscr{P}}}$. Recalling condition ($a_1$), we have that, for any given $\varepsilon>0$, there exists a point $x_\varepsilon^*\in\mathscr{E}$ such that:

$$F(x_\varepsilon^*;\varepsilon)=\min_{x\in\mathscr{E}}F(x;\varepsilon).$$

We prove, by contradiction, that there exists an $\varepsilon^*>0$ such that, for all $\varepsilon\in(0,\varepsilon^*]$ the point x_ε^* is a global solution to problem $(\tilde{\mathrm{P}})$. Suppose that this assertion is false. Then, for any integer k there must exist an $\varepsilon_k\leqq 1/k$ and a global minimizer x_k of $F(x;\varepsilon_k)$ on $\mathscr{E}$ such that x_k is not a global solution of problem $(\tilde{\mathrm{P}})$. Let $\tilde{x}$ be a global minimizer of problem $(\tilde{\mathrm{P}})$; then, by (a_3) we have $\tilde{x}\in\mathscr{K}$, so that, by definition of the set $\mathscr{K}$, we have:

$$F(\tilde{x};\varepsilon_k)=f(\tilde{x}).$$

Then, as $\mathscr{K}\subseteq\mathscr{F}\cap\mathring{\mathscr{D}}\subseteq\mathscr{E}$, we can write:

$$F(x_k,\varepsilon_k)=\min_{x\in\mathscr{E}}F(x;\varepsilon_k)\leqq F(\tilde{x};\varepsilon_k)=f(\tilde{x}). \tag{2}$$

Since $\mathcal{D}$ is compact and $\mathcal{E} \subseteq \mathcal{D}$, there exists a convergent subsequence, which we relabel $\{x_k\}$, such that $\lim_{k\to\infty} x_k = \hat{x} \in \mathcal{D}$. By (2) we have:

$$\limsup_{k\to\infty} F(x_k, \varepsilon_k) \leqq f(\tilde{x}),$$

so that (a_2) implies:

$$\hat{x} \in \mathcal{F} \cap \mathcal{D} \quad \text{and} \quad f(\hat{x}) \leqq f(\tilde{x}),$$

whence it follows that $\hat{x}$ is a global minimum point of problem $(\tilde{P})$. Recalling Assumption (A1), we have $\hat{x} \in \mathring{\mathcal{D}}$ and therefore, since $\lim_{k\to\infty} x_k = \hat{x}$, it follows that for sufficiently large values of k, say $k \geqq k_0$, the point x_k belongs to $\mathring{\mathcal{D}}$. As $\mathring{\mathcal{D}} \subseteq \mathcal{E}$, this implies that

$$F(x_k, \varepsilon_k) = \min_{x\in\mathcal{E}} F(x; \varepsilon_k) \leqq \min_{x\in\mathring{\mathcal{D}}} F(x; \varepsilon_k),$$

that is: $\mathcal{G}_{\mathcal{D}}(\varepsilon_k) \neq \emptyset$ and $x_k \in \mathcal{G}_{\mathcal{D}}(\varepsilon_k)$ for $k \geqq k_0$. Moreover, since $\hat{x} \in \mathcal{G}_{\tilde{\mathcal{P}}}$, $\lim_{k\to\infty} \varepsilon_k = 0$ and $\lim_{k\to\infty} x_k = \hat{x}$, we have that there exists an integer $k_1 \geqq k_0$ such that, for all $k \geqq k_1$, $x_k \in \mathcal{G}_{\mathcal{D}}(\varepsilon_k)$, $\varepsilon_k \leqq \varepsilon(\hat{x})$ and $\|x_k - \hat{x}\| \leqq \sigma(\hat{x})$, where $\varepsilon(\hat{x})$ and $\sigma(\hat{x})$ are the numbers considered in condition (a_4). Therefore (a_4) implies that $x_k \in \mathcal{K}$ for all $k \geqq k_1$, so that, by (2), we obtain:

$$f(x_k) = F(x_k; \varepsilon_k) \leqq f(\tilde{x}), \qquad k \geqq k_1.$$

Hence, $x_k \in \mathcal{F} \cap \mathring{\mathcal{D}}$ is both a global minimum point of $F(x; \varepsilon_k)$ on $\mathcal{E}$ and a global minimum point of problem $(\tilde{P})$ and this contradicts our original assumption. It can be concluded that there exists an $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*]$ any global minimizer x_ε^* of $F(x; \varepsilon)$ on $\mathcal{E}$ is a global solution to problem $(\tilde{P})$. On the other hand, by Assumption (A1), the global solutions of problem $(\tilde{P})$ are in $\mathring{\mathcal{D}}$ and hence, for all $\varepsilon \in (0, \varepsilon^*]$, we have that $x_\varepsilon^* \in \mathring{\mathcal{D}}$ is a global minimizer for problem (Q).

Thus we have proved that for $\varepsilon \in (0, \varepsilon^*]$, $\mathcal{G}_{\mathcal{D}}(\varepsilon) \neq \emptyset$ and $\mathcal{G}_{\mathcal{D}}(\varepsilon) \subseteq \mathcal{G}_{\tilde{\mathcal{P}}}$. Now let $\varepsilon \in (0, \varepsilon^*]$ and let x_ε be any point in $\mathcal{G}_{\mathcal{D}}(\varepsilon) \subseteq \mathcal{G}_{\tilde{\mathcal{P}}}$. By condition (a_3) we have $\mathcal{G}_{\mathcal{D}}(\varepsilon) \subseteq \mathcal{K}$ so that:

$$f(x_\varepsilon) = F(x_\varepsilon; \varepsilon). \tag{3}$$

If $\bar{x}$ is another global minimizer of Problem $(\tilde{P})$, again by (a_3), we have

$$f(\bar{x}) = F(\bar{x}; \varepsilon). \tag{4}$$

Therefore, as $f(x_\varepsilon) = f(\bar{x})$, (3) and (4) imply that $F(\bar{x}; \varepsilon) = F(x_\varepsilon; \varepsilon)$ and this proves that $\bar{x}$ is a global solution to problem (Q). Thus, $\mathcal{G}_{\tilde{\mathcal{P}}} \subseteq \mathcal{G}_{\mathcal{D}}(\varepsilon)$ for all $\varepsilon \in (0, \varepsilon^*]$ and this completes the proof. □

A short discussion of conditions (a_1)-(a_4) is in order.

Condition (a_1) requires the existence of a global minimizer of $F(x; \varepsilon)$ on the set $\mathcal{E}$. Two cases are of interest: the case $\mathcal{E} = \mathcal{D}$ and the case $\mathcal{E} = \mathring{\mathcal{D}}$. When $\mathcal{E} = \mathcal{D}$, recalling that $\mathcal{D}$ is compact, the existence of a global minimizer is ensured by the continuity of $F(x; \varepsilon)$; if $\mathcal{E} = \mathring{\mathcal{D}}$, we must specify some further condition on the behavior of $F(x; \varepsilon)$ on $\partial\mathcal{D}$. We shall address this problem later in connection with sufficient conditions for global exactness.

Condition (a_2) indicates the role played by the penalty parameter ε; it requires, in particular, that, as ε goes to zero, if the penalty function remains bounded from above, the constraints are satisfied in the limit.

With regard to (a_3), we may note that this condition is satisfied whenever, for all $\varepsilon > 0$:

$$F(x; \varepsilon) = f(x), \quad \text{for } x \in \mathcal{F} \cap \mathring{\mathcal{D}}.$$

In fact, in this case we have, by (1), that $\mathscr{K} = \mathscr{F} \cap \mathring{\mathscr{D}}$. The different classes of exact penalty functions considered in the sequel are associated with different characterizations of $\mathscr{K}$. In particular, in the case of nondifferentiable penalty functions, we have $\mathscr{K} = \mathscr{F} \cap \mathring{\mathscr{D}}$. However, this requirement would be too strong to allow the construction of continuously differentiable exact penalty functions, as will be apparent from the content of § 5. Thus, in the case of continuously differentiable exact penalty functions, the set $\mathscr{K}$ turns out to be a subset of $\mathscr{F} \cap \mathring{\mathscr{D}}$ containing a region where suitable necessary optimality conditions for problem $(\tilde{\mathrm{P}})$ are satisfied.

Finally, we may note from the proof of Theorem 1 that condition (a_4) is of major relevance in order to establish the properties of exactness considered, since the first three conditions are usually satisfied in the case of sequential penalty functions also.

We give now a sufficient condition for $F(x; \varepsilon)$ to be an exact penalty function in the sense of Definition 2, which is obtained by replacing (a_4) of Theorem 1 with a stronger condition and by imposing that $F(x; \varepsilon)$ is bounded above by $f(x)$ on the set $\mathscr{F} \cap \mathring{\mathscr{D}}$.

More specifically, we state the following theorem.

THEOREM 2. *Let $F(x; \varepsilon)$ be such that conditions (a_1)-(a_3) of Theorem 1 are satisfied and assume further that the following conditions hold.*

(a_5) *There exists an $\bar{\varepsilon} > 0$ such that, for all $\varepsilon \in (0, \bar{\varepsilon}]$, if $\mathscr{L}_{\mathscr{D}}(\varepsilon) \neq \emptyset$ and $x_\varepsilon \in \mathscr{L}_{\mathscr{D}}(\varepsilon)$, we have $x_\varepsilon \in \mathscr{K}$;*

(a_6) *$F(x; \varepsilon) \leqq f(x)$ for all $\varepsilon > 0$ and $x \in \mathscr{F} \cap \mathring{\mathscr{D}}$.*

Then, $F(x; \varepsilon)$ is an exact penalty function for problem (P) *with respect to the set $\mathscr{D}$, in the sense of Definition* 2.

Proof. We observe first that condition (a_5) is stronger than condition (a_4) of Theorem 1 so that the function $F(x; \varepsilon)$ is a weakly exact penalty function in the sense of Definition 1. Hence, there exists an $\tilde{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \tilde{\varepsilon}]$ we have $\mathscr{G}_{\mathscr{D}}(\varepsilon) = \mathscr{G}_{\mathscr{F}}$ and this implies $\mathscr{L}_{\mathscr{D}}(\varepsilon) \neq \emptyset$. Now let $\varepsilon^* = \min[\tilde{\varepsilon}, \bar{\varepsilon}]$, where $\bar{\varepsilon}$ is the number considered in condition (a_5), let $\varepsilon \in (0, \varepsilon^*]$, and assume that $x_\varepsilon \in \mathscr{L}_{\mathscr{D}}(\varepsilon)$. Then, by (a_5), we have $x_\varepsilon \in \mathscr{K}$ so that we get $F(x_\varepsilon; \varepsilon) = f(x_\varepsilon)$. This implies that for $\varepsilon \in (0, \varepsilon^*]$ and for some $\rho > 0$ it can be written:

$$f(x_\varepsilon) \leqq F(x; \varepsilon) \quad \text{for all } x \in \mathscr{F} \cap \mathscr{B}(x_\varepsilon; \rho). \tag{5}$$

Hence, by (5) and (a_6) we have:

$$f(x_\varepsilon) \leqq f(x) \quad \text{for all } x \in \mathscr{F} \cap \mathscr{B}(x_\varepsilon; \rho),$$

so that x_ε is a local minimizer of problem (P). □

As already observed in the proof of the preceding theorem, condition (a_5) is considerably stronger than (a_4) of Theorem 1; it requires, in particular, that for sufficiently small values of ε *any local minimum point of problem* (Q) *is a feasible point for problem* (P). It will be shown in the sequel that satisfaction of condition (a_5) requires the introduction of a suitable constraint qualification in problem (P).

In order to give sufficient conditions for strong exactness, we now establish a condition which ensures that isolated compact sets of local minimizers of problem (P) correspond to local unconstrained minimizers of $F(x; \varepsilon)$ for sufficiently small values of ε.

PROPOSITION 2. *Let $\mathscr{C}^*$ be a nonempty isolated compact set of local minimum points of problem* (P) *corresponding to the local minimum value f^*, such that $\mathscr{C}^* \subset \mathscr{F} \cap \mathring{\mathscr{D}}$. Let $F(x; \varepsilon)$ be such that the following conditions are satisfied.*

(b_1) *If $\{\varepsilon_k\}$ and $\{x_k\} \subset \mathring{\mathscr{D}}$ are sequences such that $\lim_{k\to\infty} \varepsilon_k = 0$, $\lim_{k\to\infty} x_k = \hat{x} \in \mathscr{D}$, and $\limsup_{k\to\infty} F(x_k, \varepsilon_k) < \infty$, we have $\hat{x} \in \mathscr{F} \cap \mathscr{D}$ and $f(\hat{x}) \leqq \limsup_{k\to\infty} F(x_k, \varepsilon_k)$.*

(b_2) $\mathscr{C}^* \subseteq \mathscr{K}$.

(b_3) *For any* $\tilde{x} \in \mathscr{C}^*$ *there exist numbers* $\varepsilon(\tilde{x}) > 0$ *and* $\sigma(\tilde{x}) > 0$ *such that, for all* $\varepsilon \in (0, \varepsilon(\tilde{x})]$, *if* $\mathscr{L}_{\mathscr{D}}(\varepsilon) \neq \emptyset$ *and* $x_\varepsilon \in \mathscr{L}_{\mathscr{D}}(\varepsilon)$ *is a local minimum point of problem* (Q) *satisfying* $\|x_\varepsilon - \tilde{x}\| \leqq \sigma(\tilde{x})$, *we have* $x_\varepsilon \in \mathscr{K}$.

Then, there exists an $\varepsilon^* > 0$ *such that for all* $\varepsilon \in (0, \varepsilon^*]$, $\bar{x} \in \mathscr{C}^*$ *implies that* $\bar{x}$ *is a local minimum point of problem* (Q).

Proof. Let $\mathscr{K}$ be the compact set considered in Lemma 1. Since $\mathscr{C}^* \subset \mathring{\mathscr{D}} \cap \mathring{\mathscr{K}}$, we can find a compact set $\mathscr{R} \subset \mathbb{R}^n$ satisfying

$$\mathscr{C}^* \subset \mathring{\mathscr{R}} \quad \text{and} \quad \mathscr{R} \subset \mathring{\mathscr{D}} \cap \mathring{\mathscr{K}},$$

such that

$$f(x) > f^* \quad \text{for all } x \in \mathscr{F} \cap \mathscr{R}, \qquad x \notin \mathscr{C}^*. \tag{6}$$

Now consider the following problem.

$$\text{minimize } f(x), \qquad x \in \mathscr{F} \cap \mathscr{R}. \tag{7}$$

Then, by (6), we have that $\mathscr{C}^* \subset \mathring{\mathscr{R}}$ is the set of global solutions of problem (7). Recalling Theorem 1, it can be easily verified that the function $F(x; \varepsilon)$ is weakly exact with respect to the set $\mathscr{R}$, and hence there exists an $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*]$, $\bar{x} \in \mathscr{C}^*$ implies that $\bar{x}$ is a global minimum point of the following problem.

$$\text{minimize } F(x; \varepsilon), \qquad x \in \mathring{\mathscr{R}}. \tag{8}$$

On the other hand, since $\mathring{\mathscr{R}} \subset \mathring{\mathscr{D}}$, any global solution of problem (8) is a local minimizer of problem (Q) and this completes the proof. □

Using the preceding result, we can establish a sufficient condition for strong exactness under the following assumption on problem (P).

Assumption (A2). There exists a finite number of isolated compact sets $\mathscr{C}^*(f_i^*)$, $i = 1, \cdots, r$ of local minimum points of problem (P) corresponding to the local minimum values f_i^*, such that $\mathscr{C}^*(f_i^*) \subset \mathring{\mathscr{D}}$ and

$$\mathscr{L}_{\mathscr{P}} \cap \mathring{\mathscr{D}} = \bigcup_{i=1}^{r} \mathscr{C}^*(f_i^*).$$

Assumption (A2) requires that any local minimizer of problem (P) in $\mathring{\mathscr{D}}$ belongs to an isolated compact set of local minimizers and that the number of these sets contained in $\mathring{\mathscr{D}}$ is finite.

Then, we have the following theorem.

THEOREM 3. *Suppose that Assumption* (A2) *holds. Let* $F(x; \varepsilon)$ *be such that the following conditions are satisfied.*

(c_1) *For any* $\varepsilon > 0$, *the function* $F(x; \varepsilon)$ *admits a global minimum point on a set* $\mathscr{E}$, *such that* $\mathring{\mathscr{D}} \subseteq \mathscr{E} \subseteq \mathscr{D}$.

(c_2) *If* $\{\varepsilon_k\}$ *and* $\{x_k\} \subset \mathscr{E}$ *are sequences such that* $\lim_{k\to\infty} \varepsilon_k = 0$, $\lim_{k\to\infty} x_k = \hat{x} \in \mathscr{D}$, *and* $\limsup_{k\to\infty} F(x_k, \varepsilon_k) < \infty$, *we have* $\hat{x} \in \mathscr{F} \cap \mathscr{D}$ *and* $f(\hat{x}) \leqq \limsup_{k\to\infty} F(x_k, \varepsilon_k)$.

(c_3) $\mathscr{L}_{\mathscr{P}} \cap \mathring{\mathscr{D}} \subseteq \mathscr{K}$.

(c_4) *There exists an* $\bar{\varepsilon} > 0$ *such that, for all* $\varepsilon \in (0, \bar{\varepsilon}]$, *if* $\mathscr{L}_{\mathscr{D}}(\varepsilon) \neq \emptyset$ *and* $x_\varepsilon \in \mathscr{L}_{\mathscr{D}}(\varepsilon)$, *we have* $x_\varepsilon \in \mathscr{K}$;

(c_5) $F(x; \varepsilon) \leqq f(x)$ *for all* $\varepsilon > 0$ *and* $x \in \mathscr{F} \cap \mathring{\mathscr{D}}$.

Then, $F(x; \varepsilon)$ *is a strongly exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$.

Proof. By Assumption (A1) we have $\mathscr{G}_{\mathscr{P}} \subseteq \mathscr{L}_{\mathscr{P}} \cap \mathring{\mathscr{D}}$; then, noting that the set of conditions (c_1)-(c_5) implies the conditions stated in Theorem 2, we have that the

function $F(x;\varepsilon)$ is exact in the sense of Definition 2 for some threshold value $\varepsilon_0>0$ of the penalty parameter. On the other hand, for $i=1,\cdots,r$ we have $\mathscr{C}^*(f_i^*)\subseteq\mathscr{K}$, so that, since conditions (c_1)-(c_5) also imply conditions (b_1)-(b_3) of Proposition 2, there exist values $\varepsilon_i>0$, $i=1,\cdots,r$ such that for all $\varepsilon\in(0,\varepsilon_i]$ we can assert that $\mathscr{C}^*(f_i^*)\subseteq\mathscr{L}_{\mathscr{D}}(\varepsilon)$. Thus, by letting $\varepsilon^*=\min_{0\leqq i\leqq r}\varepsilon_i$ we have that $F(x;\varepsilon)$ is exact, and, moreover, by Assumption (A2) we have

$$\mathscr{L}_{\mathscr{D}}\cap\mathring{\mathscr{D}}=\bigcup_{i=1}^{r}\mathscr{C}^*(f_i^*)\subseteq\mathscr{L}_{\mathscr{D}}(\varepsilon)$$

for all $\varepsilon\in(0,\varepsilon^*]$, so that the function $F(x;\varepsilon)$ is strongly exact in the sense of Definition 3. □

5. Nondifferentiable exact penalty functions. In this section we shall make use of the sufficient conditions given before in order to establish the exactness of the best-known class of nondifferentiable penalty functions.

We suppose that Assumption (A1) stated in § 4 is satisfied, that is, $\mathscr{G}_{\mathscr{D}}\subset\mathring{\mathscr{D}}$.

We consider the class of nondifferentiable penalty functions defined by

$$J_q(x;\varepsilon):=f(x)+\frac{1}{\varepsilon}\|[g^+(x)'h(x)']'\|_q,$$

where $1\leqq q\leqq\infty$. In particular, we have

$$J_q(x;\varepsilon)=f(x)+\frac{1}{\varepsilon}\left[\sum_{i=1}^{m}(g_i^+(x))^q+\sum_{j=1}^{p}|h_j(x)|^q\right]^{1/q},$$

for $1\leqq q<\infty$, and

$$J_\infty(x;\varepsilon)=f(x)+\frac{1}{\varepsilon}\max[g_1^+(x),\cdots,g_m^+(x),|h_1(x)|,\cdots,|h_p(x)|].$$

For this class of functions the set $\mathscr{K}$ defined in (1) is obviously obtained as

$$\mathscr{K}=\mathscr{F}\cap\mathring{\mathscr{D}}.$$

The expression of the directional derivative $DJ_q(x,d;\varepsilon)$ of $J_q(x;\varepsilon)$ is given in the following proposition, which is proved in full detail in [16].

PROPOSITION 3. *For all $\varepsilon>0$ and $d\in\mathbb{R}^n$, the function $J_q(x;\varepsilon)$ admits a directional derivative $DJ_q(x,d;\varepsilon)$.*

Let

$$\xi_i(x,d):=\begin{cases}\nabla g_i(x)'d, & \text{if } g_i(x)>0;\\ (\nabla g_i(x)'d)^+, & \text{if } g_i(x)=0;\\ 0, & \text{if } g_i(x)<0\end{cases}\tag{9}$$

and

$$\zeta_j(x,d):=\begin{cases}\nabla h_j(x)'d, & \text{if } h_j(x)>0;\\ |\nabla h_j(x)'d|, & \text{if } h_j(x)=0;\\ -\nabla h_j(x)'d, & \text{if } h_j(x)<0.\end{cases}\tag{10}$$

Then, we have:

(a) *for $q=1$*:

$$DJ_1(x,d;\varepsilon)=\nabla f(x)'d+\frac{1}{\varepsilon}\left(\sum_{i=1}^{m}\xi_i(x,d)+\sum_{j=1}^{p}\zeta_j(x,d)\right);$$

(b) *for* $1<q<\infty$, $x\notin\mathscr{F}$:

$$DJ_q(x,d;\varepsilon)=\nabla f(x)'d+\frac{1}{\varepsilon\|[g^+(x)'h(x)']'\|_q^{q-1}}\left[\sum_{i=1}^{m}(g_i^+(x))^{q-1}\xi_i(x,d)+\sum_{j=1}^{p}|h_j(x)|^{q-1}\zeta_j(x,d)\right];$$

(c) *for* $1<q<\infty$, $x\in\mathscr{F}$:

$$DJ_q(x,d;\varepsilon)=\nabla f(x)'d+\frac{1}{\varepsilon}\left[\sum_{i=1}^{m}(\xi_i(x,d))^q+\sum_{j=1}^{p}(\zeta_j(x,d))^q\right]^{1/q};$$

(d) *for* $q=\infty$:

$$DJ_\infty(x,d;\varepsilon)=\nabla f(x)'d+\frac{1}{\varepsilon}\max\left[\{\xi_i(x,d),i\in I_1(x)\},\{\zeta_j(x,d),j\in I_2(x)\}\right],$$

where

$$I_1(x):=\{i:\ g_i^+(x)=\|[g^+(x)'h(x)']'\|_\infty\}$$
$$I_2(x):=\{j:\ |h_j(x)|=\|[g^+(x)'h(x)']'\|_\infty\}.$$

The next two propositions, which have been proved in [17], play a significant role in establishing the exactness properties of $J_q(x;\varepsilon)$.

PROPOSITION 4. *Let* $\hat{x}\in\mathscr{F}$ *and assume that the* MFCQ *holds at* $\hat{x}$. *Then, there exist numbers* $\varepsilon(\hat{x})>0$ *and* $\sigma(\hat{x})>0$ *such that, for all* $\varepsilon\in(0,\varepsilon(\hat{x})]$, *if* x_ε *is a critical point of* $J_q(x;\varepsilon)$ *satisfying* $\|x_\varepsilon-\hat{x}\|\leqq\sigma(\hat{x})$, *we have* $x_\varepsilon\in\mathscr{F}$.

PROPOSITION 5. *Assume that the* EMFCQ *holds on* $\mathscr{D}$. *Then, there exists an* $\bar{\varepsilon}>0$ *such that, for all* $\varepsilon\in(0,\bar{\varepsilon}]$, *if* $x_\varepsilon\in\mathscr{D}$ *is a critical point of* $J_q(x;\varepsilon)$, *we have* $x_\varepsilon\in\mathscr{F}$.

We can now prove that, under suitable assumptions on problem (P), the function $J_q(x;\varepsilon)$ satisfies the sufficient conditions of exactness stated in the preceding section.

THEOREM 4. (a) *Assume that the* MFCQ *is satisfied at every global minimum point of problem* $(\tilde{\mathrm{P}})$. *Then, the function* $J_q(x;\varepsilon)$ *is a weakly exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$.

(b) *Assume that the* EMFCQ *is satisfied on* $\mathscr{D}$. *Then, the function* $J_q(x;\varepsilon)$ *is an exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$; *moreover, if Assumption* (A2) *holds, the function* $J_q(x;\varepsilon)$ *is a strongly exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$.

Proof. We show first that conditions (a_1)-(a_4) of Theorem 1 are satisfied.

Let $\mathscr{E}=\mathscr{D}$; then, (a_1) follows from the continuity of $J_q(x;\varepsilon)$ and the compactness of $\mathscr{D}$.

With regard to condition (a_2), let $\{\varepsilon_k\}$ and $\{x_k\}\subset\mathscr{D}$ be sequences such that $\varepsilon_k>0$, $\lim_{k\to\infty}\varepsilon_k=0$, $\lim_{k\to\infty}x_k=\hat{x}\in\mathscr{D}$, and assume that

$$\limsup_{k\to\infty}J_q(x_k;\varepsilon_k)=\eta<\infty.$$

This implies

$$f(\hat{x})+\limsup_{k\to\infty}\frac{1}{\varepsilon_k}\|[g^+(x_k)'h(x_k)']'\|_q=\eta,$$

whence it follows that

$$\|[g^+(\hat{x})'h(\hat{x})']'\|_q=0\quad\text{and}\quad f(\hat{x})\leqq\eta$$

so that (a_2) is satisfied.

Condition (a_3) follows from the definition of $J_q(x;\varepsilon)$, since

$$J_q(x;\varepsilon)=f(x) \quad \text{for all } x\in\mathcal{F}.$$

Finally, condition (a_4) follows from Proposition 4, since any global minimum point of problem (Q) is a critical point of $J_q(x;\varepsilon)$. This concludes the proof of (a).

Consider now the conditions stated in Theorem 2; we have already proved that (a_1)-(a_3) hold. Condition (a_5) is implied by the result given in Proposition 5 and condition (a_6) follows from the definition of $J_q(x;\varepsilon)$. It can be easily verified also that conditions (c_1), (c_2), (c_4), and (c_5) of Theorem 3 reduce to conditions (a_1), (a_2), (a_5), and (a_6), and that condition (c_3) follows from the definition of $J_q(x;\varepsilon)$. Thus (b) follows from Theorems 2 and 3. □

In the next two propositions we report additional results concerning the correspondence between critical points of $J_q(x;\varepsilon)$ and K-T triples of problem (P).

PROPOSITION 6. *Let $\bar{x}\in\mathcal{F}$; then, if $\bar{x}$ is a critical point of $J_q(x;\varepsilon)$ we have $\bar{x}\in\mathcal{T}$. Moreover, if the* EMFCQ *holds on $\mathcal{D}$, there exists an $\varepsilon^*>0$ such that for all $\varepsilon\in(0,\varepsilon^*]$, if $x_\varepsilon\in\mathring{\mathcal{D}}$ is a critical point of $J_q(x;\varepsilon)$, we have $x_\varepsilon\in\mathcal{T}$.*

Proof. Let us define the following set.

$$\mathcal{Z}:=\{z\in\mathbb{R}^n:\nabla g_i(\bar{x})'z\leqq 0,\ i\in I_0(\bar{x}),$$

$$\nabla h_j(\bar{x})'z=0, j=1,\cdots,p,\nabla f(\bar{x})'z<0\}.$$

It is known that, by Farkas' lemma, $\mathcal{Z}=\varnothing$ implies that there exist $\bar{\lambda}\in\mathbb{R}^m$ and $\bar{\mu}\in\mathbb{R}^p$ such that $(\bar{x},\bar{\lambda},\bar{\mu})$ is a K-T triple for problem (P). (See, e.g., [23, p. 18].)

We prove first that if $\bar{x}\in\mathcal{F}$ is a critical point of $J_q(x;\varepsilon)$, we have $\mathcal{Z}=\varnothing$. In fact, since $\bar{x}\in\mathcal{F}$, we have, by Proposition 3

$$DJ_q(\bar{x},z;\varepsilon)=\nabla f(\bar{x})'z+\frac{1}{\varepsilon}\left[\sum_{i\in I_0(\bar{x})}[(\nabla g_i(\bar{x})'z)^+]^q+\sum_{j=1}^{p}|\nabla h_j(\bar{x})'z|^q\right]^{1/q},$$

for $1\leqq q<\infty$, and

$$DJ_\infty(\bar{x},z;\varepsilon)=\nabla f(\bar{x})'z+\frac{1}{\varepsilon}\max[(\nabla g_i(\bar{x})'z)^+, i\in I_0(\bar{x}),|\nabla h_j(\bar{x})'z|, j=1,\cdots,p].$$

It follows that, whenever $\nabla g_i(\bar{x})'z\leqq 0$, for $i\in I_0(\bar{x})$ and $\nabla h_j(\bar{x})'z=0$, for $j=1,\cdots,p$ we have:

$$DJ_q(\bar{x},z;\varepsilon)=\nabla f(\bar{x})'z.$$

Therefore, as $\bar{x}$ is a critical point of $J_q(x;\varepsilon)$, we have $\nabla f(\bar{x})'z\geqq 0$ for all z satisfying

$$\nabla g_i(\bar{x})'z\leqq 0, \qquad i\in I_0(\bar{x})$$

$$\nabla h_j(\bar{x})'z=0, \qquad j=1,\cdots,p$$

and this implies $\mathcal{Z}=\varnothing$, so that $\bar{x}\in\mathcal{T}$. Now, recalling Proposition 5, we have that if the EMFCQ holds on $\mathcal{D}$ there exists an $\varepsilon^*>0$ such that for all $\varepsilon\in(0,\varepsilon^*]$ if $x_\varepsilon\in\mathring{\mathcal{D}}$ is a critical point of $J_q(x;\varepsilon)$ we have $x_\varepsilon\in\mathcal{F}$ and hence $x_\varepsilon\in\mathcal{T}$. □

PROPOSITION 7. *Let $(\bar{x},\bar{\lambda},\bar{\mu})$ be a* K-T *triple for Problem* (P). *Then*: (a) *$\bar{x}$ is a critical point of $J_q(x;\varepsilon)$, $1\leqq q<\infty$ for all $\varepsilon>0$ such that*

$$\bar{\lambda}_i\varepsilon\leqq(m+p)^{(1-q)/q}, \qquad i\in I_0(\bar{x})$$

$$|\bar{\mu}_j|\varepsilon\leqq(m+p)^{(1-q)/q}, \qquad j=1,\cdots,p.$$

(b) *$\bar{x}$ is a critical point of $J_\infty(x;\varepsilon)$, for all $\varepsilon>0$ such that*

$$\varepsilon\left[\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i+\sum_{j=1}^{p}|\bar{\mu}_j|\right]\leqq 1.$$

Proof. Since $(\bar{x},\bar{\lambda},\bar{\mu})$ is a K-T triple for problem (P), we can write

$$\nabla f(x)=-\left(\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i\nabla g_i(\bar{x})+\sum_{j=1}^{p}\bar{\mu}_j\nabla h_j(\bar{x})\right). \tag{11}$$

Consider first the case $1\leqq q<\infty$. As $\bar{x}\in\mathcal{F}$, by Proposition 3 we have, for any given $d\in\mathbb{R}^n$:

$$DJ_q(\bar{x},d;\varepsilon)=\nabla f(\bar{x})'d+\frac{1}{\varepsilon}\left[\sum_{i\in I_0(\bar{x})}[(\nabla g_i(\bar{x})'d)^+]^q+\sum_{j=1}^{p}|\nabla h_j(\bar{x})'d|^q\right]^{1/q},$$

so that, by (11):

$$DJ_q(\bar{x},d;\varepsilon)\geqq\frac{1}{\varepsilon}\left[\sum_{i\in I_0(\bar{x})}[(\nabla g_i(\bar{x})'d)^+]^q+\sum_{j=1}^{p}|\nabla h_j(\bar{x})'d|^q\right]^{1/q}$$
$$-\left[\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i(\nabla g_i(\bar{x})'d)^+\sum_{j=1}^{p}|\bar{\mu}_j\|\nabla h_j(\bar{x})'d|\right].$$

Using Hölder's inequality, we can write

$$DJ_q(\bar{x},d;\varepsilon)\geqq\sum_{i\in I_0(\bar{x})}\left[\frac{1}{\varepsilon(m+p)^{(q-1)/q}}-\bar{\lambda}_i\right](\nabla g_i(\bar{x})'d)^+$$
$$+\sum_{j=1}^{p}\left[\frac{1}{\varepsilon(m+p)^{(q-1)/q}}-|\bar{\mu}_j|\right]|\nabla h_j(\bar{x})'d|$$

and this implies (a).

Consider now the case $q=\infty$. Since $\bar{x}\in\mathcal{F}$, we have, by Proposition 3:

$$DJ_\infty(\bar{x},d;\varepsilon)=\nabla f(\bar{x})'d+\frac{1}{\varepsilon}\max\left[(\nabla g_i(\bar{x})'d)^+, i\in I_0(\bar{x}), |\nabla h_j(\bar{x})'d|, j=1,\cdots,p\right].$$

By (11), we can write:

$$DJ_\infty(\bar{x},d;\varepsilon)\geqq\frac{1}{\varepsilon}\max\left[(\nabla g_i(\bar{x})'d)^+, i\in I_0(\bar{x}), |\nabla h_j(\bar{x})'d|, j=1,\cdots,p\right]$$
$$-\left[\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i(\nabla g_i(\bar{x})'d)^+\sum_{j=1}^{p}|\bar{\mu}_j|\,|\nabla h_j(\bar{x})'d|\right]$$

whence, noting that

$$\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i(\nabla g_i(\bar{x})'d)^+\sum_{j=1}^{p}|\bar{\mu}_j|\,|\nabla h_j(\bar{x})'d|$$
$$\leqq\left[\sum_{i\in I_0(\bar{x})}\bar{\lambda}_i+\sum_{j=1}^{p}|\bar{\mu}_j|\right]\max\left[(\nabla g_i(\bar{x})'d)^+, i\in I_0(\bar{x}), |\nabla h_j(\bar{x})'d|, j=1,\cdots,p\right]$$

we obtain (b). □

We consider now, under suitable compactness assumptions on the feasible set, the nondifferentiable penalty function with global exactness properties proposed in [16]. This function incorporates a barrier term which goes to infinity on the boundary

of a compact perturbation of the feasible set and can be viewed as a generalization of the "M_2" penalty function introduced in [22].

Let

$$\mathcal{S}_\beta := \{x \in \mathbb{R}^n,\ g(x) \leqq \alpha,\ \|h(x)\|_2^2 \leqq \alpha_0\}$$

be the set introduced in § 2 and suppose that the following assumption is satisfied.

Assumption (A3). The set $\mathcal{S}_\beta$ is compact.

Obviously, this assumption implies that the feasible set is compact. We can take $\mathcal{D} = \mathcal{S}_\beta$, so that $\mathcal{F} \subset \mathring{\mathcal{D}}$, problem ($\tilde{\mathrm{P}}$) reduces to the original problem (P), and Assumption (A1) of § 4 is satisfied.

Let us introduce the functions:

$$a_0(x) := \alpha_0 - \|h(x)\|_2^2$$

$$a_i(x) := \alpha_i - g_i(x), \qquad i = 1, \cdots, m$$

and denote by $A(x)$ the diagonal matrix:

$$A(x) := \operatorname{diag}(a_i(x)), \qquad i = 1, \cdots, m.$$

We have, obviously, that $a_i(x) > 0$, $i = 0, 1, \cdots, m$, for all $x \in \mathring{\mathcal{D}}$.

Then, we consider the following function

$$Z_q(x; \varepsilon) := f(x) + \frac{1}{\varepsilon} \left\| \left[(A^{-1}(x) g^+(z))' \frac{h(x)'}{a_0(x)} \right]' \right\|_q,$$

where $\varepsilon > 0$ and $1 \leqq q \leqq \infty$. In particular, we have:

$$Z_q(x; \varepsilon) = f(x) + \frac{1}{\varepsilon} \left[\sum_{i=1}^{m} \left(\frac{g_i^+(x)}{a_i(x)} \right)^q + \frac{1}{a_0(x)^q} \sum_{j=1}^{p} |h_j(x)|^q \right]^{1/q},$$

for $1 \leqq q < \infty$, and

$$Z_\infty(x; \varepsilon) = f(x) + \frac{1}{\varepsilon} \max \left[\frac{g_1^+(x)}{a_1(x)}, \cdots, \frac{g_m^+(x)}{a_m(x)}, \frac{|h_1(x)|}{a_0(x)}, \cdots, \frac{|h_p(x)|}{a_0(x)} \right].$$

An equivalent expression of $Z_q(x; \varepsilon)$ can be derived by defining the functions

$$\hat{g}_i(x) = \frac{g_i(x)}{a_i(x)}, \qquad i = 1, \cdots, m \tag{12}$$

$$\hat{h}_j(x) = \frac{h_j(x)}{a_0(x)}, \qquad j = 1, \cdots, p. \tag{13}$$

Using (12) and (13) we can write

$$Z_q(x; \varepsilon) = f(x) + \frac{1}{\varepsilon} \|[\hat{g}^+(x)' \hat{h}(x)']'\|_q.$$

Taking this into account, the expression of the directional derivative $DZ_q(x, d; \varepsilon)$ can be obtained from (a)-(d) of Proposition 4 by replacing $g(x)$ with $\hat{g}(x)$ and $h(x)$ with $\hat{h}(x)$. By this substitution we have for the gradients

$$\nabla \hat{g}_i(x) = \frac{a_i}{a_i^2(x)} \nabla g_i(x), \qquad i = 1, \cdots, m \tag{14}$$

$$\nabla \hat{h}_j(x) = \frac{1}{a_0(x)} \nabla h_j(x) + 2 \frac{h_j(x)}{a_0^2(x)} \frac{\partial h(x)'}{\partial x} h(x), \qquad j = 1, \cdots, p. \tag{15}$$

We can now perform the analysis of the exactness properties of the function $Z_q(x; \varepsilon)$ making use of the sufficient conditions given in the preceding section. The

analysis is based on the fact that, by construction, the function $Z_q(x;\varepsilon)$ is defined for all $x \in \mathring{\mathcal{D}}$ and goes to infinity for x converging to a point of $\partial\mathcal{D}$; then, by Definition 4 we have that $Z_q(x;\varepsilon)$ is a globally (weakly, strongly) exact penalty function for problem (P) with respect to the set $\mathcal{D}$ if it is (weakly, strongly) exact in the sense of Definitions 1, 2, and 3.

The following propositions, which have been established in [16], can be viewed as the analogues of Propositions 5 and 6.

PROPOSITION 8. *Let $\hat{x} \in \mathcal{F}$ and assume that the* MFCQ *holds at $\hat{x}$. Then, there exist numbers $\varepsilon(\hat{x}) > 0$ and $\sigma(\hat{x}) > 0$ such that, for all $\varepsilon \in (0, \varepsilon(\hat{x})]$, if $x_\varepsilon \in \mathring{\mathcal{D}}$ is a critical point of $Z_q(x;\varepsilon)$ satisfying $\|x_\varepsilon - \hat{x}\| \leqq \sigma(\hat{x})$, we have $x_\varepsilon \in \mathcal{F}$.*

PROPOSITION 9. *Assume that the* EMFCQ *holds on $\mathcal{D}$. Then, there exists an $\bar{\varepsilon} > 0$ such that, for all $\varepsilon \in (0, \bar{\varepsilon}]$, if $x_\varepsilon \in \mathring{\mathcal{D}}$ is a critical point of $Z_q(x;\varepsilon)$, we have $x_\varepsilon \in \mathcal{F}$.*

Using the preceding results, it is possible to establish the properties of exactness of $Z_q(x;\varepsilon)$, which are collected in the following theorem.

THEOREM 5. (a) *Assume that the* MFCQ *is satisfied at every global minimum point of problem* (P). *Then, the function $Z_q(x;\varepsilon)$ is a globally weakly exact penalty function for problem* (P) *with respect to the set $\mathcal{D}$.*

(b) *Assume that the* EMFCQ *is satisfied on $\mathcal{D}$. Then, the function $Z_q(x;\varepsilon)$ is a globally exact penalty function for problem* (P) *with respect to the set $\mathcal{D}$; moreover, if Assumption* (A2) *holds, the function $Z_q(x;\varepsilon)$ is a globally strongly exact penalty function for problem* (P) *with respect to the set $\mathcal{D}$.*

Proof. By construction, we have $\lim_{k\to\infty} Z_q(x_k;\varepsilon) = \infty$ for any sequence $\{x_k\} \subset \mathring{\mathcal{D}}$ such that $x_k \to y \in \partial\mathcal{D}$. Hence, by Definition 4 we have that $Z_q(x;\varepsilon)$ is globally (weakly, strongly) exact if it is (weakly, strongly) exact.

With regard to 5(a), letting $\mathcal{C} = \mathring{\mathcal{D}}$, we can proceed as in the proof of Theorem 4 making use of Proposition 8 in place of Proposition 4; the proof of 5(b) is similar to that of Theorem 4(b) provided that we employ Proposition 9 in place of Proposition 5. □

Finally, we state without proof the relationships between critical points of $Z_q(x;\varepsilon)$ and K-T triples of problem (P).

Noting that, for $x \in \mathcal{F}$ we have $\nabla \hat{g}_i(x)'z \leqq 0$ and $\nabla \hat{h}_j(x)'z = 0$ if and only if $\nabla g_i(x)'z \leqq 0$ and $\nabla h_j(x)'z = 0$, and recalling Proposition 9, the proof of Proposition 6 can be easily modified to yield the following result.

PROPOSITION 10. *Let $\bar{x} \in \mathcal{F}$; then, if $\bar{x}$ is a critical point of $Z_q(x;\varepsilon)$, we have $\bar{x} \in \mathcal{T}$. Moreover, if the* EMFCQ *holds on $\mathcal{D}$, there exists an $\varepsilon^* > 0$ such that for all $\varepsilon \in (0, \varepsilon^*]$, if $x_\varepsilon \in \mathring{\mathcal{D}}$ is a critical point of $Z_q(x;\varepsilon)$, we have $x_\varepsilon \in \mathcal{T}$.*

The next proposition is an analogue of Proposition 7 which can be established by taking into account formulas (14) and (15) when considering the expression $DZ_q(x, d;\varepsilon)$.

PROPOSITION 11. *Let $(\bar{x}, \bar{\lambda}, \bar{\mu})$ be a K-T triple for problem* (P). *Then*: (a) *$\bar{x}$ is a critical point of $Z_q(x;\varepsilon)$, $1 \leqq q < \infty$ for all $\varepsilon > 0$ such that*:

$$\bar{\lambda}_i \varepsilon \leqq \frac{1}{\alpha_i}(m+p)^{(1-q)/q}, \qquad i \in I_0(\bar{x})$$

$$|\bar{\mu}_j|\varepsilon \leqq \frac{1}{\alpha_0}(m+p)^{(1-q)/q}, \qquad j = 1, \cdots, p.$$

(b) *$\bar{x}$ is a critical point of $Z_\infty(x;\varepsilon)$, for all $\varepsilon > 0$ such that*:

$$\varepsilon\left[\sum_{i\in I_0(\bar{x})} \bar{\lambda}_i \alpha_i + \sum_{j=1}^{p} |\bar{\mu}_j|\alpha_0\right] \leqq 1.$$

6. Continuously differentiable exact penalty functions. In this section we study a class of continuously differentiable exact penalty functions, making use of the sufficient conditions established in § 4.

The key idea for the construction of continuously differentiable exact penalty functions is that of replacing the multiplier vectors (λ, μ) which appear in the augmented Lagrangian function of Hestenes, Powell, and Rockafellar [4] with continuously differentiable *multiplier functions* $(\lambda(x), \mu(x))$, depending on the problem variables.

Let

$$\mathscr{X} := \{x \in \mathbb{R}^n : \nabla g_i(x),\ i \in I_0(x),\ \nabla h_j(x),\ j = 1, \cdots, p \text{ are linearly independent}\};$$

then, for any $x \in \mathscr{X}$ we can consider the multiplier functions $(\lambda(x), \mu(x))$ introduced in [28], which are obtained by minimizing over $\mathbb{R}^m \times \mathbb{R}^p$ the quadratic function in (λ, μ) defined by:

$$\Psi(\lambda, \mu; x) := \|\nabla_x L(x, \lambda, \mu)\|^2 + \gamma^2 \|G(x)\lambda\|^2,$$

where $\gamma \neq 0$ and

$$G(x) := \operatorname{diag}(g_i(x)).$$

The function $\Psi(\lambda, \mu; x)$ can be viewed as a measure of the violation of the set of K-T necessary conditions:

$$\nabla_x L(x, \lambda, \mu) = 0, \qquad G(x)\lambda = 0.$$

Let $N(x)$ be the $(m+p)\times(m+p)$ matrix defined by:

$$N(x) := \begin{bmatrix} \dfrac{\partial g(x)}{\partial x}\dfrac{\partial g(x)'}{\partial x} + \gamma^2 G^2(x) & \dfrac{\partial g(x)}{\partial x}\dfrac{\partial h(x)'}{\partial x} \\ \dfrac{\partial h(x)}{\partial x}\dfrac{\partial g(x)'}{\partial x} & \dfrac{\partial h(x)}{\partial x}\dfrac{\partial h(x)'}{\partial x} \end{bmatrix}.$$

In the next proposition we recall some known results established in [28].

PROPOSITION 12. *Let $\bar{x} \in \mathscr{X}$ and $\gamma \neq 0$. Then*: (a) *the matrix $N(x)$ is positive definite*;

(b) *there exists a unique minimizer $(\lambda(x), \mu(x))$ of the quadratic function in (λ, μ), $\Psi(\lambda, \mu; x)$, given by*

$$\begin{bmatrix} \lambda(x) \\ \mu(x) \end{bmatrix} = -N^{-1}(x) \begin{bmatrix} \dfrac{\partial g(x)}{\partial x} \\ \dfrac{\partial h(x)}{\partial x} \end{bmatrix} \nabla f(x);$$

(c) *if $(\bar{x}, \bar{\lambda}, \bar{\mu}) \in \mathscr{X} \times \mathbb{R}^m \times \mathbb{R}^p$ is a triple such that $\nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) = 0$ and $G(\bar{x})\bar{\lambda} = 0$, we have $\lambda(\bar{x}) = \bar{\lambda}$ and $\mu(\bar{x}) = \bar{\mu}$*;

(d) *the Jacobian matrices of $\lambda(x)$ and $\mu(x)$ are given by*

$$\begin{bmatrix} \dfrac{\partial \lambda(x)}{\partial x} \\ \dfrac{\partial \mu(x)}{\partial x} \end{bmatrix} = -N^{-1}(x) \begin{bmatrix} R(x) \\ S(x) \end{bmatrix}, \tag{16}$$

where

$$R(x) := \frac{\partial g(x)}{\partial x} \nabla_x^2 L(x, \lambda(x), \mu(x)) + \sum_{i=1}^{m} e_i^m \nabla_x L(x, \lambda(x), \mu(x))' \nabla^2 g_i(x)$$

$$+ 2\gamma^2 \Lambda(x) G(x) \frac{\partial g(x)}{\partial x}$$

$$S(x) := \frac{\partial h(x)}{\partial x} \nabla_x^2 L(x, \lambda(x), \mu(x)) + \sum_{j=1}^{p} e_j^p \nabla_x L(x, \lambda(x), \mu(x))' \nabla^2 h_j(x)$$

$$\nabla_x L(x, \lambda(x), \mu(x)) := [\nabla_x L(x, \lambda, \mu)]_{\substack{\lambda=\lambda(x) \\ \mu=\mu(x)}}$$

$$\nabla_x^2 L(x, \lambda(x), \mu(x)) := [\nabla_x^2 L(x, \lambda, \mu)]_{\substack{\lambda=\lambda(x) \\ \mu=\mu(x)}}$$

$$\Lambda(x) := \operatorname{diag}(\lambda_i(x))$$

and $e_i^m (e_j^p)$ *denote the* ith$(j$th$)$ *column of the* $m \times m (p \times p)$ *identity matrix.*

Thus we can consider the penalty function introduced in [28], defined by

$$W(x; \varepsilon) := f(x) + \lambda(x)'(g(x) + Y(x; \varepsilon) y(x; \varepsilon)) + \frac{1}{\varepsilon} \|g(x) + Y(x; \varepsilon) y(x; \varepsilon)\|^2 + \mu(x)' h(x) + \frac{1}{\varepsilon} \|h(x)\|^2, \tag{17}$$

where

$$y_i(x; \varepsilon) := \left\{ -\min \left[0, g_i(x) + \frac{\varepsilon}{2} \lambda_i(x) \right] \right\}^{1/2}, \qquad i = 1, \cdots, m$$
$$Y(x; \varepsilon) := \operatorname{diag}(y_i(x; \varepsilon)). \tag{18}$$

It can be verified that the function $W(x; \varepsilon)$ can also be written in the form

$$W(x; \varepsilon) = f(x) + \lambda(x)' g(x) + \frac{1}{\varepsilon} \|g(x)\|^2 + \mu(x)' h(x) + \frac{1}{\varepsilon} \|h(x)\|^2 - \frac{1}{4\varepsilon} \sum_{i=1}^{m} \{\min [0, \varepsilon \lambda_i(x) + 2 g_i(x)]\}^2.$$

From the above expression and the differentiability assumptions on the problem functions, it follows that $W(x; \varepsilon)$ is continuously differentiable on $\mathscr{X}$. The expression of $W(x; \varepsilon)$ can be derived by means of the following reasoning.

Consider the transformed problem:

$$\text{minimize } f(x)$$
$$\text{subject to } g(x) + Yy = 0, \qquad h(x) = 0,$$

where y_i, $i = 1, \cdots, m$ are slack variables and $Y := \operatorname{diag}(y_i)$.

Define the augmented Lagrangian function for this problem:

$$L_a(x, y, \lambda, \mu; \varepsilon) := f(x) + \lambda'(g(x) + Yy) + \frac{1}{\varepsilon} \|g(x) + Yy\|^2 + \mu' h(x) + \frac{1}{\varepsilon} \|h(x)\|^2.$$

Then, by substituting $(\lambda(x), \mu(x))$ for (λ, μ) and minimizing with respect to y, we get the function $W(x; \varepsilon)$, that is,

$$W(x; \varepsilon) = L_a(x, y(x; \varepsilon), \lambda(x), \mu(x); \varepsilon) = \min_y L_a(x, y, \lambda(x), \mu(x); \varepsilon).$$

Since, by construction,

$$[\nabla_y L_a(x, y, \lambda, \mu; \varepsilon)]_{\substack{\lambda=\lambda(x) \\ \mu=\mu(x) \\ y=y(x; \varepsilon)}} = 0,$$

the gradient expression of $W(x;\varepsilon)$ can be obtained by treating formally $y(x;\varepsilon)$ as a constant vector. Thus, we have:

$$\nabla W(x;\varepsilon)=\nabla f(x)+\frac{\partial g(x)'}{\partial x}\lambda(x)+\frac{\partial h(x)'}{\partial x}\mu(x)$$
$$+\frac{2}{\varepsilon}\frac{\partial g(x)'}{\partial x}(g(x)+Y(x;\varepsilon)y(x;\varepsilon))+\frac{2}{\varepsilon}\frac{\partial h(x)'}{\partial x}h(x)$$
$$+\frac{\partial \lambda(x)'}{\partial x}(g(x)+Y(x;\varepsilon)y(x;\varepsilon))+\frac{\partial \mu(x)'}{\partial x}h(x),$$

where $\partial\lambda(x)/\partial x$ and $\partial\mu(x)/\partial x$ are the Jacobian matrices defined in (16).

Some properties of exactness of the function $W(x;\varepsilon)$ have been established in [15] for inequality constrained problems. Here we perform a more complete analysis for problems with both equality and inequality constraints, making use of the sufficient conditions given in § 4.

We suppose that Assumption (A1) of § 4 is satisfied, that is, $\mathscr{G}_{\mathscr{D}}\subset\mathring{\mathscr{D}}$, and that everywhere in this section the following assumption holds.

Assumption (A4). The LICQ is satisfied on $\mathscr{D}$, that is, $\mathscr{D}\subset\mathscr{K}$.

Some immediate consequences of the definition of $W(x;\varepsilon)$ are pointed out in the following proposition.

PROPOSITION 13. *Let* $(\bar{x},\bar{\lambda},\bar{\mu})$ *be a* K-T *triple for problem* (P), *such that* $\bar{x}\in\mathscr{D}$. *Then, for any* $\varepsilon>0$, *we have*: (a) $g(\bar{x})+Y(\bar{x};\varepsilon)y(\bar{x};\varepsilon)=0$;

(b) $W(\bar{x};\varepsilon)=f(\bar{x})$;

(c) $\nabla W(\bar{x};\varepsilon)=0$.

Proof. By Proposition 12 we have $\lambda(\bar{x})=\bar{\lambda}$ and $\mu(\bar{x})=\bar{\mu}$, so that, since $(\bar{x},\bar{\lambda},\bar{\mu})$ is a K-T triple for problem (P), we obtain $\nabla_x L(\bar{x},\lambda(\bar{x}),\mu(\bar{x}))=0$, $\lambda(\bar{x})\geqq 0$ and $\lambda_i(\bar{x})=0$ when $g_i(\bar{x})<0$. Then, (a) is satisfied by definition of $y(\bar{x};\varepsilon)$ and (b) follows directly from (17). Finally, (c) follows from (19), taking (a) into account and noting that, by assumption, $h(\bar{x})=0$ and $\nabla_x L(\bar{x},\lambda(\bar{x}),\mu(\bar{x}))=\nabla_x L(\bar{x},\bar{\lambda},\bar{\mu})$. □

Then, we have the following proposition.

PROPOSITION 14. *Let* $\hat{x}\in\mathscr{F}\cap\mathscr{D}$. *Then, there exist numbers* $\varepsilon(\hat{x})>0$ *and* $\sigma(\hat{x})>0$ *such that, for all* $\varepsilon\in(0,\varepsilon(\hat{x})]$, *if* $x_\varepsilon\in\mathring{\mathscr{D}}$ *is a stationary point of* $W(x;\varepsilon)$ *satisfying* $\|x_\varepsilon-\hat{x}\|\leqq\sigma(\hat{x})$, *we have that* $(x_\varepsilon,\lambda(x_\varepsilon),\mu(x_\varepsilon))$ *is a* K-T *triple for problem* (P).

Proof. Let $x\in\mathscr{K}$; then, by definition of $y(x;\varepsilon)$, we have

$$Y^2(x;\varepsilon)\lambda(x)=-\frac{2}{\varepsilon}Y^2(x;\varepsilon)(g(x)+Y(x;\varepsilon)y(x;\varepsilon)); \tag{20}$$

moreover, by definition of $\lambda(x)$ we can write:

$$\frac{\partial g(x)}{\partial x}\nabla_x L(x,\lambda(x),\mu(x))=-\gamma^2G^2(x)\lambda(x)$$
$$=-\gamma^2G(x)(G(x)+Y^2(x;\varepsilon))\lambda(x)+\gamma^2G(x)Y^2(x;\varepsilon)\lambda(x) \tag{21}$$
$$=-\gamma^2G(x)\Lambda(x)(g(x)+Y(x;\varepsilon)y(x;\varepsilon))+\gamma^2G(x)Y^2(x;\varepsilon)\lambda(x).$$

Therefore, by (20) and (21) we get

$$\varepsilon\frac{\partial g(x)}{\partial x}\nabla_x L(x,\lambda(x),\mu(x))=-\gamma^2G(x)(\varepsilon\Lambda(x)+2Y^2(x;\varepsilon))(g(x)+Y(x;\varepsilon)y(x;\varepsilon)),$$

so that, by (19), we can write

$$
\begin{aligned}
\varepsilon \frac{\partial g(x)}{\partial x} \nabla W(x;\varepsilon) &= \varepsilon \frac{\partial g(x)}{\partial x} \nabla_x L(x, \lambda(x), \mu(x)) \\
&\quad + \frac{\partial g(x)}{\partial x}\left(2\frac{\partial g(x)'}{\partial x} + \varepsilon \frac{\partial \lambda(x)'}{\partial x}\right)(g(x) + Y(x;\varepsilon)y(x;\varepsilon)) \\
&\quad + \frac{\partial g(x)}{\partial x}\left(2\frac{\partial h(x)'}{\partial x} + \varepsilon \frac{\partial \mu(x)'}{\partial x}\right)h(x) \\
&= K_{11}(x;\varepsilon)(g(x) + Y(x;\varepsilon)y(x;\varepsilon)) + K_{12}(x;\varepsilon)h(x),
\end{aligned} \tag{22}
$$

where

$$
\begin{aligned}
K_{11}(x;\varepsilon) &:= 2\left(\frac{\partial g(x)}{\partial x}\frac{\partial g(x)'}{\partial x} - \gamma^2 G(x) Y^2(x;\varepsilon)\right) \\
&\quad + \varepsilon\left(\frac{\partial g(x)}{\partial x}\frac{\partial \lambda(x)'}{\partial x} - \gamma^2 G(x)\Lambda(x)\right), \\
K_{12}(x;\varepsilon) &:= 2\frac{\partial g(x)}{\partial x}\frac{\partial h(x)'}{\partial x} + \varepsilon \frac{\partial g(x)}{\partial x}\frac{\partial \mu(x)'}{\partial x}.
\end{aligned}
$$

Now, by definition of $\mu(x)$, we have

$$
\frac{\partial h(x)}{\partial x} \nabla_x L(x, \lambda(x), \mu(x)) = 0
$$

and hence, by (19) we can write

$$
\varepsilon \frac{\partial h(x)}{\partial x} \nabla W(x;\varepsilon) = K_{21}(x;\varepsilon)(g(x) + Y(x;\varepsilon)y(x;\varepsilon)) + K_{22}(x;\varepsilon)h(x), \tag{23}
$$

where

$$
\begin{aligned}
K_{21}(x;\varepsilon) &:= 2\frac{\partial h(x)}{\partial x}\frac{\partial g(x)'}{\partial x} + \varepsilon \frac{\partial h(x)}{\partial x}\frac{\partial \lambda(x)'}{\partial x} \\
K_{22}(x;\varepsilon) &:= 2\frac{\partial h(x)}{\partial x}\frac{\partial h(x)'}{\partial x} + \varepsilon \frac{\partial h(x)}{\partial x}\frac{\partial \mu(x)'}{\partial x}.
\end{aligned}
$$

Thus, from (22) and (23) we get, for all $x \in \mathscr{X}$:

$$
\varepsilon \begin{bmatrix} \dfrac{\partial g(x)}{\partial x} \\ \dfrac{\partial h(x)}{\partial x} \end{bmatrix} \nabla W(x;\varepsilon) = K(x;\varepsilon) \begin{bmatrix} g(x) + Y(x;\varepsilon)y(x;\varepsilon) \\ h(x) \end{bmatrix}, \tag{24}
$$

where $K(x;\varepsilon)$ is the matrix defined by

$$
K(x;\varepsilon) := \begin{bmatrix} K_{11}(x;\varepsilon) & K_{12}(x;\varepsilon) \\ K_{21}(x;\varepsilon) & K_{22}(x;\varepsilon) \end{bmatrix}.
$$

Let now $\hat{x} \in \mathscr{F} \cap \mathscr{D}$; then, by definition of $y(x;\varepsilon)$ we have

$$
Y^2(\hat{x};0) = -G(\hat{x}),
$$

so that, by definition of $K(x;\varepsilon)$, we get

$$
K(\hat{x};0) = 2N(\hat{x}).
$$

Therefore, since Assumption (A4) implies that $N(x)$ is nonsingular, by continuity there exist numbers $\varepsilon(\hat{x})>0$ and $\sigma(\hat{x})>0$ such that the matrix $K(x;\varepsilon)$ is nonsingular for all $\varepsilon\in[0,\varepsilon(\hat{x})]$ and all x such that $\|x-\hat{x}\|\leqq\sigma(\hat{x})$. Let $\varepsilon\in[0,\varepsilon(\hat{x})]$ and let $x_\varepsilon\in\mathring{\mathscr{D}}$ be a stationary point of $W(x;\varepsilon)$ satisfying $\|x_\varepsilon-\hat{x}\|\leqq\sigma(\hat{x})$. By (24) we have

$$K(x_\varepsilon;\varepsilon)\begin{bmatrix} g(x_\varepsilon)+Y(x_\varepsilon;\varepsilon)y(x_\varepsilon;\varepsilon) \\ h(x_\varepsilon)\end{bmatrix}=0,$$

which implies, as $K(x_\varepsilon;\varepsilon)$ is nonsingular, $h(x_\varepsilon)=0$, and

$$g(x_\varepsilon)+Y(x_\varepsilon;\varepsilon)y(x_\varepsilon;\varepsilon)=0. \tag{25}$$

Therefore, since $\nabla W(x_\varepsilon;\varepsilon)=0$, we have from (19)

$$\nabla_x L(x_\varepsilon,\lambda(x_\varepsilon),\mu(x_\varepsilon))=0; \tag{26}$$

on the other hand, by definition of $\lambda(x)$ and $\mu(x)$, we have

$$\frac{\partial g(x)}{\partial x}\nabla_x L(x_\varepsilon,\lambda(x_\varepsilon),\mu(x_\varepsilon))+\gamma^2 G^2(x_\varepsilon)\lambda(x_\varepsilon)=0, \tag{27}$$

and hence, by (26) and (27) we obtain

$$G(x_\varepsilon)\lambda(x_\varepsilon)=0.$$

Finally, if $g_i(x_\varepsilon)=0$ for some i, we have, by (25), $y_i^2(x_\varepsilon;\varepsilon)=0$, which implies, by definition of $y(x;\varepsilon)$, that $\lambda_i(x_\varepsilon)\geqq 0$. Hence, the triple $(x_\varepsilon,\lambda(x_\varepsilon),\mu(x_\varepsilon))$ is a K-T triple for problem (P). □

The next proposition establishes the correspondence between stationary points of $W(x;\varepsilon)$ and K-T triples for problem (P) on the whole set $\mathscr{D}$.

PROPOSITION 15. *Assume that the* EMFCQ *holds on* $\mathscr{D}$. *Then, there exists an* $\varepsilon^*>0$ *such that, for all* $\varepsilon\in(0,\varepsilon^*]$, *if* $x_\varepsilon\in\mathring{\mathscr{D}}$ *is a stationary point of* $W(x;\varepsilon)$, *we have that* $(x_\varepsilon,\lambda(x_\varepsilon),\mu(x_\varepsilon))$ *is a* K-T *triple for problem* (P).

Proof. The proof is by contradiction. Assume that the assertion is false. Then, for any integer k, there exists an $\varepsilon_k\leqq 1/k$ and a point $x_k\in\mathring{\mathscr{D}}$ such that $\nabla W(x_k;\varepsilon_k)=0$, but $(x_k,\lambda(x_k),\mu(x_k))$ is not a K-T triple for problem (P).

Since $\mathscr{D}$ is compact, there exists a convergent subsequence (relabel it again $\{x_k\}$) such that $\lim_{k\to\infty}x_k=\hat{x}\in\mathscr{D}$. Moreover, since $\nabla W(x_k;\varepsilon_k)=0$ for all k and since $\varepsilon_k\to 0$, we have in the limit, by (19):

$$\frac{\partial g(\hat{x})'}{\partial x}(g(\hat{x})+Y(\hat{x};0)y(\hat{x};0))+\frac{\partial h(\hat{x})'}{\partial x}h(\hat{x})=0, \tag{28}$$

where, by definition of $y_i^2(x;\varepsilon)$ we have

$$y_i^2(\hat{x};0)=-\min[0,g_i(\hat{x})],\qquad i=1,\cdots,m.$$

It follows that (28) can be rewritten into the form:

$$\sum_{i\in I_+(\hat{x})}g_i(\hat{x})\nabla g_i(\hat{x})+\sum_{j=1}^{p}h_j(\hat{x})\nabla h_j(\hat{x})=0,$$

where $I_+(\hat{x})=\{i\colon g_i(\hat{x})\geqq 0\}$. Therefore, by the EMFCQ we have $g_i(\hat{x})=0$, $i\in I_+(\hat{x})$, and $h_j(\hat{x})=0, j=1,\cdots,p$ so that $\hat{x}\in\mathscr{F}\cap\mathring{\mathscr{D}}$. On the other hand, by Proposition 14 there exists an integer $\bar{k}$ such that for all $k\geqq\bar{k}$ we have that $(x_k,\lambda(x_k),\mu(x_k))$ is a K-T triple for problem (P) and we get a contradiction. □

The properties of exactness of $W(x;\varepsilon)$ are summarized in the following theorem.

THEOREM 6. (a) *The function* $W(x;\varepsilon)$ *is a weakly exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$.

(b) *Assume that the* EMFCQ *is satisfied on* $\mathscr{D}$. *Then, the function* $W(x;\varepsilon)$ *is an exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$; *moreover, if Assumption* (A2) *holds, the function* $W(x;\varepsilon)$ *is a strongly exact penalty function for problem* (P) *with respect to the set* $\mathscr{D}$.

Proof. Let $\mathscr{E}=\mathscr{D}$; we show first that conditions (a_1)-(a_4) of Theorem 1 are satisfied.

It is easily seen that (a_1) follows from the continuity of $W(x;\varepsilon)$ and the compactness of $\mathscr{D}$.

With regard to condition (a_2), let $\{\varepsilon_k\}$ and $\{x_k\}\subset\mathscr{D}$ be sequences such that $\varepsilon_k>0$, $\lim_{k\to\infty}\varepsilon_k=0$, $\lim_{k\to\infty}x_k=\hat{x}\in\mathscr{D}$ and assume that

$$\limsup_{k\to\infty} W(x_k;\varepsilon_k)<\infty.$$

By the continuity assumptions we get from (17)

$$g(\hat{x})+Y(\hat{x};0)y(\hat{x};0)=0, \qquad h(\hat{x})=0,$$

which imply $\hat{x}\in\mathscr{F}$ and

$$f(\hat{x})\leqq\limsup_{k\to\infty} W(x_k;\varepsilon_k),$$

so that (a_2) is satisfied.

We observe now that Assumptions (A1) and (A4) imply that $\mathscr{G}_{\mathscr{F}}\subseteq\mathscr{T}$. On the other hand, by (b) of Proposition 13 we obtain $\mathscr{T}\subseteq\mathscr{K}$. Therefore, we have $\mathscr{G}_{\mathscr{F}}\subseteq\mathscr{K}$ and hence condition (a_3) is satisfied.

Finally, condition (a_4) follows from Proposition 14, noting that $\mathscr{T}\subseteq\mathscr{K}$ and that, by the differentiability of $W(x;\varepsilon)$, any point $x_\varepsilon\in\mathscr{G}_{\mathscr{D}}(\varepsilon)$ is a stationary point of $W(x;\varepsilon)$. Thus (a) is proved.

With regard to (b), we have already shown that conditions (a_1)-(a_3) of Theorem 2 are satisfied. Condition (a_5) is implied by Proposition 15. Therefore we must show that (a_6) holds, that is,

$$W(x;\varepsilon)\leqq f(x) \quad \text{for all } \varepsilon>0 \quad \text{and} \quad x\in\mathscr{F}\cap\mathring{\mathscr{D}}.$$

Let $x\in\mathring{\mathscr{D}}$ be such that $g(x)\leqq 0$ and $h(x)=0$. Suppose first that $y_i^2(x;\varepsilon)=0$; this implies, by definition of $y_i^2(x;\varepsilon)$, that

$$2g_i(x)+\varepsilon\lambda_i(x)\geqq 0,$$

so that, since $g_i(x)\leqq 0$, we have

$$\frac{1}{\varepsilon}g_i^2(x)+\lambda_i(x)g_i(x)\leqq\frac{2}{\varepsilon}g_i^2(x)+\lambda_i(x)g_i(x)\leqq 0. \tag{29}$$

Now assume that $y_i^2(x;\varepsilon)>0$; in this case we obtain

$$g_i(x)+y_i^2(x;\varepsilon)=-\frac{\varepsilon}{2}\lambda_i(x),$$

whence:

$$\frac{1}{\varepsilon}(g_i(x)+y_i^2(x;\varepsilon))^2+\lambda_i(x)(g_i(x)+y_i^2(x;\varepsilon))=-\frac{\varepsilon}{4}\lambda_i^2(x)\leqq 0. \tag{30}$$

Therefore, by (29) and (30) we have, for any $i=1,\cdots,m$:

$$\frac{1}{\varepsilon}(g_i(x)+y_i^2(x;\varepsilon))^2+\lambda_i(x)(g_i(x)+y_i^2(x;\varepsilon))\leqq 0,$$

so that, by (17), we obtain $W(x;\varepsilon)\leqq f(x)$ and hence condition (a_6) is satisfied. It can be verified also that conditions (c_1), (c_2), (c_4), and (c_5) of Theorem 3 are satisfied and that condition (c_3) follows from (b) of Proposition 13. Thus (b) follows from Theorems 2 and 3. □

We now consider the construction of a continuously differentiable exact penalty function with global exactness properties, along the same lines followed in the non-differentiable case. As in § 5, we take $\mathcal{D}=\mathcal{S}_\beta$, and we suppose that Assumption (A3) holds. Then, we can define on the set $\mathcal{E}=\mathring{\mathcal{D}}$ the continuously differentiable exact penalty function:

$$Z(x;\varepsilon):=f(x)+\lambda(x)'(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon))$$
$$+\frac{1}{\varepsilon}(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon))'A^{-1}(x)(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon))$$
$$+\mu(x)'h(x)+\frac{1}{\varepsilon a_0(x)}\|h(x)\|^2,$$

where $(\lambda(x),\mu(x))$ are the multiplier functions defined in Proposition 12, and

$$a_0(x):=\alpha_0-\|h(x)\|_2^2$$
$$a_i(x):=\alpha_i-g_i(x),\qquad i=1,\cdots,m$$
$$A(x):=\operatorname{diag}(a_i(x)),\qquad i=1,\cdots,m$$

with:

$$\tilde{Y}(x;\varepsilon):=\operatorname{diag}(\tilde{y}_i(x;\varepsilon)),\qquad i=1,\cdots,m$$
$$\tilde{y}_i(x;\varepsilon):=\left\{-\min\left[0,\, g_i(x)+\frac{\varepsilon}{2}a_i(x)\lambda_i(x)\right]\right\}^{1/2}.$$

In order to justify the expression of $Z(x;\varepsilon)$ we first consider the equality constrained problem obtained from problem (P) by introducing the vector Yy of squared slack variables into the inequality constraints $g(x)\leqq 0$.

Then, we define the augmented Lagrangian function:

$$\tilde{L}_a(x,y,\lambda,\mu;\varepsilon):=f(x)+\lambda'(g(x)+Yy)+\frac{1}{\varepsilon}(g(x)+Yy)'A^{-1}(x)(g(x)+Yy)$$
$$+\mu'h(x)+\frac{1}{\varepsilon a_0(x)}\|h(x)\|^2,$$

where the penalty terms are weighted by the barrier functions $A^{-1}(x)$ and $1/a_0(x)$.

Finally, by substituting $(\lambda(x),\mu(x))$ for (λ,μ) and minimizing with respect to y we get the function $Z(x;\varepsilon)$, that is,

$$Z(x;\varepsilon)=\tilde{L}_a(x,\tilde{y}(x;\varepsilon),\lambda(x),\mu(x);\varepsilon)$$
$$=\min_y \tilde{L}_a(x,y,\lambda(x),\mu(x);\varepsilon).$$

By construction, we have

$$[\nabla_y\tilde{L}_a(x,y,\lambda,\mu)]_{\substack{\lambda=\lambda(x)\\ \mu=\mu(x)\\ y=\tilde{y}(x;\varepsilon)}}=0$$

and hence the gradient expression of $Z(x;\varepsilon)$ can be obtained by taking $\tilde{y}(x;\varepsilon)$ as a constant vector.

Thus, we can write:

$$\begin{aligned}\nabla Z(x;\varepsilon) = {} & \nabla f(x) + \frac{\partial g(x)'}{\partial x}\lambda(x) + \frac{\partial h(x)'}{\partial x}\mu(x) \\ & + \frac{2}{\varepsilon}\frac{\partial g(x)'}{\partial x}A^{-1}(x)(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon)) \\ & + \frac{1}{\varepsilon}\frac{\partial g(x)'}{\partial x}[G(x)+\tilde{Y}^2(x;\varepsilon)]A^{-2}(x)(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon)) \\ & + \frac{2}{\varepsilon a_0(x)}\frac{\partial h(x)'}{\partial x}h(x) + \frac{2\|h(x)\|_2^2}{a_0^2(x)}\frac{\partial h(x)'}{\partial x}h(x) \\ & + \frac{\partial \lambda(x)'}{\partial x}(g(x)+\tilde{Y}(x;\varepsilon)\tilde{y}(x;\varepsilon)) + \frac{\partial \mu(x)'}{\partial x}h(x).\end{aligned} \tag{31}$$

The study of the properties of exactness of the function $Z(x;\varepsilon)$ can be performed along the same lines followed in the case of the function $W(x;\varepsilon)$, taking into account the expressions of $\tilde{y}(x;\varepsilon)$ and $\nabla Z(x;\varepsilon)$ and noting that $a_i(x) > 0$, $i = 0, 1, \cdots, m$ for $x \in \mathring{\mathcal{D}}$.

In particular, the next two propositions are the analogue of Propositions 13 and 14 and can be proved in a similar way.

PROPOSITION 16. *Let* $(\bar{x}, \bar{\lambda}, \bar{\mu})$ *be a* K-T *triple for problem* (P), *such that* $\bar{x} \in \mathcal{D}$. *Then, for any* $\varepsilon > 0$, *we have*:

(a) $g(\bar{x}) + \tilde{Y}(\bar{x};\varepsilon)\tilde{y}(\bar{x};\varepsilon) = 0$;

(b) $Z(\bar{x};\varepsilon) = f(\bar{x})$;

(c) $\nabla Z(\bar{x};\varepsilon) = 0$.

PROPOSITION 17. *Let* $\hat{x} \in \mathcal{F} \cap \mathcal{D}$. *Then, there exist numbers* $\varepsilon(\hat{x}) > 0$ *and* $\sigma(\hat{x}) > 0$ *such that, for all* $\varepsilon \in (0, \varepsilon(\hat{x})]$, *if* $x_\varepsilon \in \mathring{\mathcal{D}}$ *is a stationary point of* $Z(x;\varepsilon)$ *satisfying* $\|x_\varepsilon - \hat{x}\| \leqq \sigma(\hat{x})$, *we have that* $(x_\varepsilon, \lambda(x_\varepsilon), \mu(x_\varepsilon))$ *is a* K-T *triple for problem* (P).

We now need the following lemma which is proved in [15].

LEMMA 3. *Let* $\{\delta_k^{(i)}\}$, $i = 1, \cdots, r$ *be* r *sequences of positive numbers. Then, there exist an index* i^* *and subsequences* $\{\delta_k^{(i)}\}_K$, $i = 1, \cdots, r$ *corresponding to the same index set* K, *such that*:

$$\lim_{\substack{k\to\infty \\ k\in K}} \frac{\delta_k^{(i^*)}}{\delta_k^{(i)}} = l_i < +\infty, \quad i = 1, \cdots, r.$$

The following proposition establishes the correspondence between stationary points of $Z(x;\varepsilon)$ and K-T triples for problem (P).

PROPOSITION 18. *Assume that the* EMFCQ *holds on* $\mathcal{D}$. *Then, there exists an* $\varepsilon^* > 0$ *such that, for all* $\varepsilon \in (0, \varepsilon^*]$, *if* $x_\varepsilon \in \mathring{\mathcal{D}}$ *is a stationary point of* $Z(x;\varepsilon)$, *we have that* $(x_\varepsilon, \lambda(x_\varepsilon), \mu(x_\varepsilon))$ *is a* K-T *triple for problem* (P).

Proof. Reasoning by contradiction, we assume that for any integer k, there exists an $\varepsilon_k \leqq 1/k$ and a point $x_k \in \mathring{\mathcal{D}}$ such that $\nabla Z(x_k;\varepsilon_k) = 0$, but $(x_k, \lambda(x_k), \mu(x_k))$ is not a K-T triple for problem (P). Since $\mathcal{D}$ is compact, there exists a convergent subsequence (relabel it again $\{x_k\}$) such that $\lim_{k\to\infty} x_k = \hat{x} \in \mathcal{D}$.

We show first that $\hat{x} \in \mathring{\mathscr{D}}$. In fact, assume that $\hat{x} \in \partial\mathscr{D}$; this implies, by definition of $\mathscr{D}$, that there exists a subset J of $\{0, 1, \cdots, m\}$ such that:

$$\lim_{k\to\infty} a_i(x_k) = 0, \qquad i \in J. \tag{32}$$

By Lemma 3 we can define an index $i^* \in J$ and a subsequence (relabel it again $\{x_k\}$) such that

$$\lim_{k\to\infty} \frac{a_{i^*}(x_k)}{a_i(x_k)} = l_i < +\infty, \qquad i \in J, \tag{33}$$

where, in particular, $l_{i^*} = 1$. Recalling (31), we can write

$$\begin{aligned} 0 &= \varepsilon_k a_{i^*}^2(x_k) \nabla Z(x_k; \varepsilon_k) \\ &= \varepsilon_k a_{i^*}^2 \Big[\nabla_x L(x_k, \lambda(x_k), \mu(x_k)) \\ &\qquad + \frac{\partial \lambda(x_k)'}{\partial x} (g(x_k) + \tilde{Y}(x_k; \varepsilon_k) y(x_k; \varepsilon_k)) + \frac{\partial \mu(x_k)'}{\partial x} h(x_k) \Big] \\ &\quad + \sum_{i=1}^{m} \frac{a_{i^*}^2(x_k)}{a_i(x_k)} \left(2 + \frac{g_i(x_k) + \tilde{y}_i^2(x_k; \varepsilon_k)}{a_i(x_k)} \right) (g_i(x_k) + \tilde{y}_i^2(x_k; \varepsilon_k)) \nabla g_i(x_k) \\ &\quad + 2 \sum_{j=1}^{p} \frac{a_{i^*}^2(x_k)}{a_0(x_k)} \left(1 + \frac{\|h(x_k)\|_2^2}{a_0(x_k)} \right) h_j(x_k) \nabla h_j(x_k). \end{aligned} \tag{34}$$

Taking limits of (34) and recalling (32) we can write

$$\sum_{i=1}^{m} v_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} u_j \nabla h_j(\hat{x}) = 0, \tag{35}$$

where, by (33)

$$v_i = \begin{cases} l_i^2 (g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0))^2, & \text{if } i \in J; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$u_j = \begin{cases} l_0^2 \|h(\hat{x})\|_2^2 h_j(\hat{x}) & \text{if } 0 \in J \\ 0, & \text{otherwise.} \end{cases}$$

Since $i \in J$ and $i \geqq 1$ imply $i \in I_+(\hat{x})$, we can rewrite (35) into the following form:

$$\sum_{i \in I_+(\hat{x})} v_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} u_j \nabla h_j(\hat{x}) = 0.$$

This implies, by the EMFCQ, that $v_i = 0$ for $i \in I_+(\hat{x})$ and $u_j = 0$, for $j = 1, \cdots, p$. On the other hand, as $l_{i^*} = 1$, we have either $h(\hat{x}) = 0$ (if $i^* = 0$) or $g_{i^*}(\hat{x}) + \tilde{y}_{i^*}^2(\hat{x}; 0) = 0$ (if $i^* \in \{1, \cdots, m\}$). In both cases we get a contradiction to (32). Then we can conclude that $\hat{x} \in \mathring{\mathscr{D}}$. Therefore, from (31) and (34), taking the limit of $\varepsilon_k \nabla Z(x_k; \varepsilon_k)$ over the subsequence converging to $\hat{x}$, we have

$$\begin{aligned} &\sum_{i=1}^{m} \left[2 \frac{g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0)}{a_i(\hat{x})} + \frac{(g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0))^2}{a_i^2(\hat{x})} \right] \nabla g_i(\hat{x}) \\ &\quad + \sum_{j=1}^{p} \left[2 \frac{h_j(\hat{x})}{a_0(\hat{x})} + \frac{\|h(\hat{x})\|_2^2}{a_0^2(\hat{x})} h_j(\hat{x}) \right] \nabla h_j(\hat{x}) = 0. \end{aligned}$$

Noting that, by definition of $\tilde{y}_i(x_k; \varepsilon_k)$, the inequality $g_i(\hat{x}) < 0$ implies $g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0) = 0$, we can write

$$\sum_{i \in I_+(\hat{x})} v_i \nabla g_i(\hat{x}) + \sum_{j=1}^{p} u_j \nabla h_j(\hat{x}) = 0,$$

where now

$$v_i := 2 \frac{g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0)}{a_i(\hat{x})} + \frac{(g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0))^2}{a_i^2(\hat{x})} \geqq 0$$

and

$$u_j := 2\left[1 + \frac{\|h(\hat{x})\|_2^2}{a_0(\hat{x})}\right] \frac{h_j(\hat{x})}{a_0(\hat{x})}.$$

Therefore, again by the EMFCQ, we have $v_i = 0, i \in I_+(\hat{x})$, and $\mu_j = 0, j = 1, \cdots, p$ which imply

$$g_i(\hat{x}) + \tilde{y}_i^2(\hat{x}; 0) = 0, \qquad i = 1, \cdots, m$$

$$h_j(\hat{x}) = 0, \qquad j = 1, \cdots, p,$$

so that $\hat{x} \in \mathscr{F}$. As $\varepsilon_k \to 0$, Proposition 17 implies that for sufficiently large values of k, the triple $(x_k, \lambda(x_k), \mu(x_k))$ is a K-T triple for problem (P) and this yields a contradiction. □

We can now summarize the properties of exactness of $Z(x; \varepsilon)$ in the following theorem.

THEOREM 7. (a) *The function $Z(x; \varepsilon)$ is a globally weakly exact penalty function for problem* (P) *with respect to the set $\mathscr{D}$.*

(b) *Assume that the* EMFCQ *is satisfied on $\mathscr{D}$. Then, the function $Z(x; \varepsilon)$ is a globally exact penalty function for problem* (P) *with respect to the set $\mathscr{D}$; moreover, if Assumption* (A2) *holds, the function $Z(x; \varepsilon)$ is a globally strongly exact penalty function for problem* (P) *with respect to the set $\mathscr{D}$.*

Proof. By construction, we have $\lim_{k \to \infty} Z(x_k; \varepsilon) = \infty$ for any sequence $\{x_k\} \subset \mathring{\mathscr{D}}$ such that $x_k \to y \in \partial \mathscr{D}$. Hence, by Definition 4 we have that $Z(x; \varepsilon)$ is globally (weakly, strongly) exact if it is (weakly, strongly) exact.

Letting $\mathscr{E} = \mathring{\mathscr{D}}$, assertion (a) can be proved along the same lines followed in the proof of Theorem 6, making use of Propositions 16 and 17 in place of Propositions 13 and 14.

With regard to (b), again letting $\mathscr{E} = \mathring{\mathscr{D}}$, we can proceed, as in the proof of Theorem 6, by employing Proposition 18 in place of Proposition 15 and making use of the inequality:

$$Z(x; \varepsilon) \leqq f(x) \quad \text{for all } x \in \mathscr{F},$$

which can be established in a way similar to that followed in the proof of Theorem 6 for the case of the function $W(x; \varepsilon)$. □

REFERENCES

[1] M. S. BAZARAA AND J. J. GOODE, *Sufficient conditions for a globally exact penalty function without convexity*, Math. Programming Stud., 19 (1982), pp. 1-15.

[2] D. P. BERTSEKAS, *Necessary and sufficient conditions for a penalty method to be exact*, Math. Programming, 9 (1975), pp. 87-99.

[3] D. P. BERTSEKAS, *Enlarging the region of convergence of Newton's method for constrained optimization*, J. Optim. Theory Appl., 36 (1982), pp. 221-252.
[4] ———, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.
[5] T. F. COLEMAN AND A. R. CONN, *Nonlinear programming via an exact penalty function method: asymptotic analysis*, Math. Programming, 24 (1982), pp. 123-136.
[6] ———, *Nonlinear programming via an exact penalty function method: global analysis*, Math. Programming, 24 (1982), pp. 137-161.
[7] A. R. CONN, *Constrained optimization using a nondifferentiable penalty function*, SIAM J. Numer. Anal., 10 (1973), pp. 760-784.
[8] ———, *Penalty function method*, M. J. D. Powell, ed., Nonlinear Optimization 1981, Academic Press, New York, 1982.
[9] ———, *Nonlinear programming, exact penalty functions and projection techniques for nonsmooth functions*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Numerical Optimization 1984, Society for Industrial and Applied Mathematics, Philadelphia, 1984, pp. 3-25.
[10] A. R. CONN AND T. PIETRZYKOWSKI, *A penalty function method converging directly to a constrained optimum*, SIAM J. Numer. Anal., 14 (1977), 348-378.
[11] G. DI PILLO AND L. GRIPPO, *A new class of augmented Lagrangians in nonlinear programming*, SIAM J. Control Optim., 17 (1979), pp. 618-628.
[12] ———, *An augmented Lagrangian for inequality constraints in nonlinear programming problems*, J. Optim. Theory Appl., 36 (1982), pp. 495-519.
[13] ———, *A class of continuously differentiable exact penalty function algorithms for nonlinear programming problems*, P. Toft-Christensen, ed., System Modelling and Optimization, Springer-Verlag, Berlin, 1984, pp. 246-256.
[14] ———, *A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints*, SIAM J. Control Optim., 23 (1985), pp. 72-84.
[15] ———, *An exact penalty method with global convergence properties for nonlinear programming problems*, Math. Programming, 36 (1986), pp. 1-18.
[16] ———, *Globally exact nondifferentiable penalty functions*, Report 10.87, (1987), Dipartimento di Informatica e Sistemistica, University of Rome "La Sapienza", Rome.
[17] ———, *On the exactness of a class of non-differentiable penalty functions*, J. Optim. Theory Appl., 57 (1988), pp. 399-410.
[18] G. DI PILLO, F. FACCHINEI, AND L. GRIPPO, *A RQP algorithm using a differentiable exact penalty function for inequality constrained problems*, Report R-127, (1988), IASI, National Research Council, Rome.
[19] G. DI PILLO, L. GRIPPO, AND F. LAMPARIELLO, *A method for solving equality constrained optimization problems by unconstrained minimization*, K. Iracki, K. Malanowski, and S. Walukiewicz, eds., Optimization Techniques, Springer-Verlag, Berlin, 1980.
[20] ———, *A class of methods for the solution of optimization problems with inequalities*, R. F. Drenick and F. Kozin, eds., System Modelling and Optimization, Springer-Verlag, Berlin, 1981.
[21] L. C. W. DIXON, *Exact penalty function methods in nonlinear programming*, Report NOC, (1979), The Hatfield Polytechnic, n. 103.
[22] J. P. EVANS, F. J. GOULD, AND J. W. TOLLE, *Exact penalty functions in nonlinear programming*, Math. Programming, 4 (1973), pp. 72-97.
[23] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.
[24] R. E. FLETCHER, *A class of methods for nonlinear programming with termination and convergence properties*, J. Abadie, ed., Integer and Nonlinear Programming, North-Holland, Amsterdam, 1970, pp. 157-173.
[25] ———, *An exact penalty function for nonlinear programming with inequalities*, Math. Programming, 5 (1973), pp. 129-150.
[26] ———, *Penalty functions*, A. Bachem, M. Grötschel, and B. Korte, eds., Mathematical Programming, The State of the Art, Springer-Verlag, Berlin, 1983, pp. 87-114.
[27] ———, *An l_1 penalty method for nonlinear constraints*, P. T. Boggs, R. H. Byrd, and R. B. Schnabel, eds., Numerical Optimization 1984, Society for Industrial and Applied Mathematics, Philadelphia, 1984, pp. 26-40.
[28] T. GLAD AND E. POLAK, *A multiplier method with automatic limitation of penalty growth*, Math. Programming, 17 (1979), pp. 140-155.
[29] S. P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251-269.
[30] R. A. HORN AND C. A. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[31] S. LUCIDI, *New results on a class of exact augmented Lagrangians*, J. Optim. Theory Appl., 58 (1988), pp. 259–282.

[32] O. L. MANGASARIAN, *Nonlinear Programming*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[33] D. Q. MAYNE AND E. POLAK, *A superlinearly convergent algorithm for constrained optimization problems*, Math. Programming Stud., 16 (1982), pp. 45–61.

[34] H. MUKAI AND E. POLAK, *A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints*, Math. Programming, 9 (1975), pp. 336–349.

[35] T. PIETRZYKOWSKI, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1969), pp. 294–304.

[36] M. J. D. POWELL, *Methods for nonlinear constraints in optimization calculations*, Report DAMTP 1986/NA5, (1986), University of Cambridge, Cambridge.

[37] M. J. D. POWELL AND Y. YUAN, *A recursive quadratic programming algorithm that uses differentiable exact penalty functions*, Math. Programming, 3 (1986), pp. 265–278.

[38] C. VINANTE AND S. PINTOS, *On differentiable exact penalty functions*, J. Optim. Theory Appl., 50 (1986), pp. 479–493.

[39] W. I. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1361-1368, November 1989

007

A GEOMETRIC ISOMORPHISM WITH APPLICATIONS TO CLOSED LOOP CONTROLS*

ROBERT B. GARDNER†, WILLIAM F. SHADWICK‡, AND GEORGE R. WILKENS§

Abstract. Feedback equivalence of n state, $n-1$ control systems satisfying certain regularity conditions divides such systems into two invariant classes. We show that class one corresponds, via a geometric isomorphism, to classical Lagrangian variational problems. We prove the existence of time critical closed loop controls for systems that satisfy the nondegeneracy condition that the analogue of the Hessian for the Lagrangian problem have full rank. We show that the vanishing of this Hessian characterizes the control linear systems in class one and identify the rank condition for local controllability for such systems as the nonvanishing of a differential invariant. The control linear systems in class two are also characterized by the vanishing of an invariant and the rank condition is identified.

Key words. feedback equivalence of control systems, classical Lagrangian variational problems, time critical closed loop controls

AMS(MOS) subject classifications. 49, 53

1. Introduction. In this paper we consider the problem of feedback equivalence of control systems, with n states and $n-1$ controls, as the equivalence problem for systems

$$\frac{dx}{dt} = F(x, u), \qquad x \in \mathbf{R}^n, \quad u \in \mathbf{R}^{n-1}, \tag{1.1}$$

under diffeomorphisms of the form

$$\Phi(t, x, u) = (t, \phi(x), \psi(x, u)). \tag{1.2}$$

By making use of Cartan's method of equivalence [3], [5], [6] we obtain an invariant splitting of regular systems into two classes. The first of these, on which we focus our attention, is identified, via a geometric isomorphism, with classical single integral variational problems. The existence of this isomorphism means that all of the rich geometry of classical Lagrangian mechanics is encoded in the control system (1.1) and may be applied to its study. The most basic elements of the Lagrangian problem are the notions of regularity and of the Euler-Lagrange equations for critical curves. We show that, as one would hope, these concepts translate into basic features of the control problem. The *vanishing* of the Hessian for the associated Lagrangian is necessary and sufficient for the control system to be equivalent to one in control-linear form

$$\frac{dx}{dt} = f(x) + \sum_{i=1}^{n-1} g_i(x) u^i. \tag{1.3}$$

At the other extreme, when the Hessian has full rank, we show that the Euler-Lagrange equations may be solved to provide closed loop controls. As the Lagrangian functional, applied to solution curves of (1.1), measures the *time* from initial to final endpoints,

* Received by the editors March 9, 1988; accepted for publication (in revised form) February 1, 1989.
† Department of Mathematics, University of North Carolina, Chapel Hill, North Carolina 27599-3250. The work of this author was supported by National Science Foundation grant DMS-8505434.
‡ Department of Mathematics, Duke University, Durham, North Carolina 27706. Permanent address, Pure Mathematics Department, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. The work of this author was supported by Natural Sciences and Engineering Research Council of Canada grant A7895.
§ Department of Mathematics, University of Hawaii at Manoa, Honolulu, Hawaii 96822.

these controls are time critical. Moreover, for a certain subclass of control systems, they are the geodesics of a pseudo-Riemannian metric intrinsic to the system as Wilkens found in the case $n=3$ [13].

The second class of systems corresponds to a nonclassical variational problem of the sort studied by Griffiths [8] and Bryant [2]. We will pursue the study of this aspect of the problem elsewhere. We show that every system in this class can be put in the form

$$\frac{dx^i}{dt}=u^i, \quad 1\leqq i\leqq n-1, \qquad \frac{dx^n}{dt}=g(x,u)$$

where g is homogeneous of degree one in the controls. We give necessary and sufficient conditions for the system to be control linear, in which case $g(x,u)=\sum g_i(x)u^i$. For such systems the controllability condition [1], [11] is that an invariant should not vanish. As there is no drift term, the condition will fail for any *linearizable* system in this class, however, there is an invariant skew symmetric matrix j that determines at least rank j equivalence classes of controllable control linear systems.[1]

The regularity mentioned after (1.2) is a set of conditions that is developed during the analysis of the problem. In particular, we will require that the rank of the $n\times(n-1)$ matrix $\partial F/\partial u$ be $n-1$ and that certain functions either vanish on an open set or never vanish on an open set. Finally we note that, while the restriction to consideration of systems with n states and $n-1$ controls is essential to the identification of a *classical first order* variational problem, the same techniques apply to the case of n states and p controls. This case requires the analysis of more general variational problems and is currently being studied.

2. The equivalence problem. Given the system (1.1) on $U_0\subset\mathbf{R}^{2n}$ with coordinates $t, x^1,\cdots,x^n, u^1,\cdots,u^{n-1}$ and a second system

$$\frac{d\bar{x}}{d\bar{t}}=\bar{F}(\bar{x},\bar{u})$$

on $\bar{U}_0\subset\mathbf{R}^{2n}$, the problem of local equivalence under feedback is the equivalence problem for maps

$$\Phi: U_0\to\bar{U}_0$$

of the form

$$\bar{t}\circ\Phi=t, \quad \bar{x}\circ\Phi=\phi(x), \quad \bar{u}\circ\Phi=\psi(x,u) \tag{2.1}$$

that satisfy

$$\Phi^*(d\bar{x}-\bar{F}\,d\bar{t})=T(dx-F\,dt) \tag{2.2}$$

for some $T: U_0\to \mathrm{GL}(n,\mathbf{R})$. This is an overdetermined equivalence problem and leads, as discussed in [7], to the following problem in standard form.

Let U and $\bar{U}$ be open sets on which F and $\bar{F}$ are nonzero and let A_0 and $\bar{A}_0$ be maps from U and $\bar{U}$ to $\mathrm{GL}(n,\mathbf{R})$ such that

$$A_0F=\bar{A}_0\bar{F}={}^t(1,0,\cdots,0).$$

A diffeomorphism Φ from U to $\bar{U}$ satisfies (2.1) and (2.2) if and only if

$$\bar{t}\circ\Phi=t$$

and

$$\Phi^*\left(\frac{\bar{A}_0\,d\bar{x}}{d\bar{u}}\right)=\left(\begin{array}{cc|c} 1 & A & 0\\ 0 & B & 0\\ \hline C & D & E\end{array}\right)\left(\frac{A_0\,dx}{du}\right) \tag{2.3}$$

[1] There is a normal form for the case where j has full rank; see *Feedback Equivalence for General Control Systems*, by Gardner and Shadwick (to appear).

with $B, E \in \mathrm{GL}(n-1, \mathbf{R})$. This now has the form considered in [5] and [6] where G is the subgroup of $\mathrm{GL}(2n-1, \mathbf{R})$ of matrices of the form

$$\begin{pmatrix} 1 & A & 0 \\ 0 & B & 0 \\ C & D & E \end{pmatrix} \tag{2.3'}$$

with B and E in $\mathrm{GL}(n-1, \mathbf{R})$. Thus we construct the vector of one-forms ${}^t(\eta^1, \cdots, \eta^n, \mu^1, \cdots, \mu^{n-1})$ on $U \times G$ given by

$$\begin{pmatrix} \eta \\ \hline \mu \end{pmatrix} = \left(\begin{array}{ccc|c} 1 & A & & 0 \\ 0 & B & & 0 \\ \hline C & D & & E \end{array}\right) \begin{pmatrix} \eta_U \\ \hline \mu_U \end{pmatrix} \tag{2.4}$$

where

$$\begin{pmatrix} \eta_U \\ \mu_U \end{pmatrix} = \begin{pmatrix} A_0\, dx \\ du \end{pmatrix}.$$

We now turn to the equivalence of classical first-order Lagrangian problems and show that it leads to the same structure. Here we consider functionals

$$\mathscr{L}(c) = \int_c L(\tau, q, \dot{q})\, d\tau$$

over curves c in $\mathbf{R}^{2p+1}$ that satisfy

$$\dot{q}^i \circ c = \frac{d}{d\tau}(q^i \circ c), \qquad 1 \leqq i \leqq p.$$

These curves are integrals of the contact system

$$\theta^i := dq^i - \dot{q}^i\, d\tau, \qquad 1 \leqq i \leqq p$$

with independence condition $d\tau \neq 0$. Given a second functional $\bar{\mathscr{L}}$, we will say that $\mathscr{L}$ and $\bar{\mathscr{L}}$ are simply equivalent [2] if there is a contact transformation Φ such that

$$\Phi^* \bar{L}\, d\bar{\tau} \equiv L\, d\tau \pmod{\theta^i}. \tag{2.5}$$

This is clearly an equivalence relation and preserves the value of the functionals on integrals of the contact system.

If we complete $\{L\, d\tau, \theta\}$ and $\{\bar{L}\, d\bar{\tau}, \bar{\theta}\}$ to coframes by adding one-forms ζ and $\bar{\zeta}$ and use the fact that Φ must preserve the contact system $\{\theta^i\}$, we may summarize the conditions on the Jacobian of an equivalence Φ by

$$\Phi^* \begin{pmatrix} \bar{L}\, d\bar{\tau} \\ \bar{\theta} \\ \bar{\zeta} \end{pmatrix} = \begin{pmatrix} 1 & A & 0 \\ 0 & B & 0 \\ C & D & E \end{pmatrix} \begin{pmatrix} L\, d\tau \\ \theta \\ \zeta \end{pmatrix}. \tag{2.6}$$

Thus, if $p = n-1$, (2.6) and (2.4) suggest the identification on η^1_U with $L\, d\tau$ and $\{\eta^2_U, \cdots, \eta^n_U\}$ with the contact system. It is easy to verify that if c is a solution of (1.1) we have

$$c^* A_0\, dx = {}^t(dt, 0, \cdots, 0)$$

so $c^* \eta^1 = dt$, $c^* \eta^i = 0$, $2 \leqq i \leqq n$, and

$$\int_c \eta^1 = \int_{t_0}^{t_1} dt.$$

Thus the variational problem we have identified is the *time optimization problem* for solutions of (1.1).

There is an integrability condition that obstructs the identification of $\{\eta^2, \cdots, \eta^n\}$ with a contact system: the derived structure of $\eta^2, \cdots, \eta^n$ must coincide with that of $\theta^1, \cdots, \theta^{n-1}$. We now show how this invariant arises by pursuing the equivalence problem calculation.

For the system (2.4) with $\bar{\eta} = {}^t(\eta^2, \cdots, \eta^n)$, the structure equations take the form

$$\begin{aligned} d\eta^1 &= {}^t\alpha \wedge \bar{\eta} + \eta^1\, {}^t m \wedge \mu, \\ d\bar{\eta} &= \beta \wedge \bar{\eta} + \eta^1 M \wedge \mu, \\ d\mu &= \gamma \wedge \eta^1 + \delta \wedge \bar{\eta} + \varepsilon \wedge \mu, \end{aligned} \tag{2.7}$$

after all torsion has been absorbed.

The matrix $\mathcal{M} := \binom{{}^t m}{M}$ is, up to left and right multiplication, just

$$A_0\left(\frac{\partial F}{\partial u}\right)$$

and hence the full rank case is the only one in which all $n-1$ controls are actually present. From the infinitesimal action on ${}^t m$ and M,

$$\begin{aligned} &d\,{}^t m - {}^t\alpha M + {}^t m\varepsilon \equiv 0 \\ &dM - \beta M + M\varepsilon \equiv 0 \end{aligned} \qquad (\mathrm{mod}\ \eta^1, \bar{\eta}, \mu), \tag{2.8}$$

we see that the rank of M is also an invariant and thus there are two cases to consider: rank $M = n-2$, and rank $M = n-1$. In either case, as we may assume that rank $\mathcal{M} = n-1$, we may put the original control system in a form similar to those observed by Hermann [9]

$$\frac{dx^i}{dt} = u^i, \quad 1 \leqq i \leqq n-1, \qquad \frac{dx^n}{dt} = g(x, u), \tag{2.9}$$

and we may choose A_0 to be given by

$$A_0 = \begin{pmatrix} 1/u^1 & 0 & 0 & \cdots & 0 \\ -u^2 & u^1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ -u^{n-1} & 0 & \cdots & u^1 & 0 \\ -g & 0 & \cdots & 0 & u^1 \end{pmatrix}. \tag{2.10}$$

A short calculation now shows that rank $M = n-2$ if and only if $\hat{g} := g - \sum u^i\, \partial g/\partial u^i = 0$, i.e., if and only if g is homogeneous of degree one, clearly a nongeneric condition.

Case 1. Rank $M = n-1$. We have $\hat{g} := g - \sum u^i\, \partial g/\partial u^i \neq 0$ and from (2.8) we may normalize M to the identity and ${}^t m$ to zero to obtain new congruences

$$\begin{aligned} &{}^t\alpha \equiv 0 \\ &\beta \equiv \varepsilon \end{aligned} \qquad (\mathrm{mod}\ \eta^1, \bar{\eta}, \mu). \tag{2.8'}$$

In this case, the identification suggested above is actually an isomorphism, as the structure equations are now identical with those of the classical Lagrangian problem [6], and we have the following theorem.

THEOREM 2.1. *If* rank $M = n-1$ *the system is isomorphic to the classical first-order Lagrangian system in* $n-1$ *dependent variables.*

The remainder of the equivalence problem is now precisely the calculations for the Lagrangian case carried out by Bryant and Gardner [6] and the parametric form of the invariants has recently been investigated by Sutton [12]. For our purposes, it suffices to consider the invariants introduced by the congruences (2.8′).

After absorption of torsion terms the structure equations become

$$\begin{aligned} d\eta^1 &= {}^t\mu H \wedge \bar{\eta} + {}^t\bar{\eta} j \wedge \bar{\eta} + \eta^1 {}^t k \wedge \bar{\eta}, \\ d\bar{\eta} &= \beta \wedge \bar{\eta} + \eta^1 \wedge \mu, \\ d\mu &= \gamma \wedge \eta^1 + \delta \wedge \bar{\eta} + \beta \wedge \mu. \end{aligned} \tag{2.11}$$

The integrability condition $d^2 = 0$ shows that H is a symmetric matrix, and in the Lagrangian variables, H is, up to conjugation, just the Hessian matrix $(\partial^2 L/\partial \dot{q}^i\, \partial \dot{q}^j)$.

THEOREM 2.2. *If $\hat{g} \neq 0$, H vanishes if and only if the system* (1.1) *is equivalent to*

$$\begin{aligned} \frac{dx^i}{dt} &= u^i, \qquad 1 \leqq i \leqq n-1, \\ \frac{dx^n}{dt} &= f(x) + \sum g_i(x) u^i. \end{aligned} \tag{2.12}$$

If $H = 0$ the rank condition for local controllability is satisfied if and only if $d\eta^1 \neq 0$. The condition $d\eta^1 = 0$ gives a conservation law generalizing Hermes [10].

Proof. To establish this result we look at the explicit parametric calculation for the normalization of ${}^t m$ to zero. If we adopt the parametrization given by (2.9) and (2.10), then

$$\begin{aligned} \eta^1 = \frac{dx^1}{u^1} &+ A_1(u^1\, dx^2 - u^2\, dx^1) + \cdots + A_{n-2}(u^1\, dx^{n-1} - u^{n-1}\, dx^1) \\ &+ A_{n-1}(u^1\, dx^n - g\, dx^1). \end{aligned}$$

It is easy to check that $d\eta^1 \equiv 0 \pmod{\bar{\eta}}$ requires

$$\begin{aligned} A_1 &= -A_{n-1} \frac{\partial g}{\partial u^2} \\ A_2 &= -A_{n-1} \frac{\partial g}{\partial u^3} \\ &\vdots \\ A_{n-2} &= -A_{n-1} \frac{\partial g}{\partial u^{n-1}} \end{aligned} \tag{2.13}$$

and $1/u^1 = A_{n-1}\hat{g}$. As $\hat{g} \neq 0$, we may solve for A_{n-1} to obtain

$$\eta^1 = \frac{1}{\hat{g}} \left(dx^n - \frac{\partial g}{\partial u^1}\, dx^1 - \frac{\partial g}{\partial u^2}\, dx^2 - \cdots - \frac{\partial g}{\partial u^{n-1}}\, dx^{n-1} \right). \tag{2.14}$$

Now $H = 0$ means

$$d\eta^1 \wedge \eta^1 = {}^t\bar{\eta} S \wedge \bar{\eta} \wedge \eta^1 \tag{2.15}$$

so the right-hand side has no component in μ and hence no component in du. But

$$\begin{aligned} d\eta^1 \wedge \eta^1 = -\frac{1}{\hat{g}^2} &\left\{ d\left(\frac{\partial g}{\partial u^1}\right) \wedge dx^1 + \cdots + d\left(\frac{\partial g}{\partial u^{n-1}}\right) \wedge dx^{n-1} \right\} \\ &\wedge \left\{ dx^n - \frac{\partial g}{\partial u^1}\, dx^1 - \frac{\partial g}{\partial u^2}\, dx^2 - \cdots - \frac{\partial g}{\partial u^{n-1}}\, dx^{n-1} \right\} \end{aligned}$$

and the vanishing of the terms in $du^i \wedge dx^n$ required by (2.15) forces $\partial^2 g/\partial u^i\, \partial u^j = 0$, $1 \leqq i, j \leqq n-1$, so $g = f(x) + \sum g_i(x) u^i$, where $f \neq 0$. It is clear that this condition is also

sufficient for $H=0$. When $H=0$ it follows from (2.14) that

$$\eta^1=\frac{1}{f(x)}\left(dx^n-g_1(x)\,dx^1-\cdots-g_{n-1}(x)\,dx^{n-1}\right). \tag{2.14'}$$

If we define $X_n:=f(x)\,\partial/\partial x^n$ and $X_i:=\partial/\partial x^i+g_i(x)\,\partial/\partial x^n$ the rank condition [1], [11] is satisfied unless $[X_n, X_i]=[X_i, X_j]=0$ for $1\leqq i, j\leqq n-1$ and it follows directly from (2.14') that this happens if and only if $d\eta^1=0$. □

Next we proceed to the case in which H has full rank. The infinitesimal action on H is given by $dH-{}^t\beta H-H\beta\equiv 0\ (\mathrm{mod}\ \eta^1, \bar{\eta}, \mu)$, showing that H is being conjugated. On an open set on which the rank and signature of H are constant, we may normalize H to a constant matrix Q with the same rank and signature. The remaining torsion terms j and k may both be normalized to zero, putting $d\eta^1$ in normal form:

$$d\eta^1={}^t\mu Q\wedge\bar{\eta}. \tag{2.16}$$

As described in [4] and more recently in [5], the *Euler-Lagrange equations* for the functional $\int\eta^1$ are the exterior equations

$$\mu=0,\qquad \bar{\eta}=0. \tag{2.17}$$

THEOREM 2.3. *If rank* $M=n-1$ *and rank* $H=n-1$ *the system* (2.17) *yields closed loop time critical controls for* (1.1).

Proof. The system $\{\eta^2,\cdots,\eta^n,\mu^1,\cdots,\mu^{n-1}\}$ is completely integrable and hence

$$\mu\equiv T\,dw\qquad(\mathrm{mod}\ \eta^2,\cdots,\eta^n)$$

for some nonsingular matrix T and vector function $w(x, u)$. Because

$$0\neq\mu^1\wedge\cdots\wedge\mu^{n-1}\wedge\eta^1\wedge\cdots\wedge\eta^n=\det T\det\frac{\partial w}{\partial u}\,du^1\wedge\cdots\wedge du^{n-1}\wedge\eta^1\wedge\cdots\wedge\eta^n$$

the system $w=z$, z constant, can be solved for $u^i(x)$, $1\leqq i\leqq n-1$. But, as we have already observed, the solutions of $dx/dt=F(x, u(x))$ solve $\bar{\eta}=0$ and, by construction, also satisfy $\mu=0$. Thus they are solutions of the Euler-Lagrange system (2.17) and as such are time critical. □

We also note that the same arguments given by Wilkens [13] show that there is a class of control systems for which the quadratic form $(\eta^1)^2+{}^t\bar{\eta}Q\bar{\eta}$ defines a pseudo-Riemannian metric on the state space and the solutions of (2.17) are geodesics of the metric.

Case 2. Rank $M=n-2$. We conclude by considering the second class of problems. After the reduction of $\mathcal{M}$ the only unabsorbable torsion is in $d\eta^n$ and

$$d\eta^n=\beta_n\wedge\eta^n+{}^t\bar{\mu}S\wedge\bar{\eta}+(\eta^1, {}^t\bar{\eta})T\wedge\begin{pmatrix}\eta^1\\ \bar{\eta}\end{pmatrix} \tag{2.18}$$

where ${}^tS=S$ and ${}^tT=-T$.

THEOREM 2.4. *If* $\hat{g}=0$ *then* $d\eta^n$ *can be put in the form* (2.18) *and* $S=0$ *if and only if the system is control linear. If* $S=0$ *the system satisfies the rank condition for local controllability if and only if* $T\neq 0$.

Proof. We can make the following choices for the one-forms η_U, μ_U:

$$\eta_U^1 = \frac{dx^1}{u^1}, \qquad \bar{\eta}_U^\alpha = dx^\alpha - u^\alpha \frac{dx^1}{u^1}, \quad 2 \leqq \alpha \leqq n-1,$$

$$\eta_U^n = dx^n - \sum \frac{\partial g}{\partial u^i}\, dx^i, \qquad \mu_U^1 = \frac{du^1}{u^1},$$

$$\bar{\mu}_U^\alpha = du^\alpha - u^\alpha \frac{du^1}{u^1}, \qquad 2 \leqq \alpha \leqq n-1,$$

and the reduction of $\mathcal{M}$ imposes the following relations on the group of matrices defined by (2.3′):

$$B = \begin{pmatrix} B_1 & B_2 \\ 0 & B_3 \end{pmatrix}, \qquad B_3 \in \mathbf{R},$$

$$E = \begin{pmatrix} 1 & A \\ 0 & B_1 \end{pmatrix}.$$

Since

$$\eta^n = B_3\left(dx^n - \sum \frac{\partial g}{\partial u^i}\, dx^i\right),$$

it is clear from (2.18) that $S = 0$ if and only if

$$\frac{\partial^2 g}{\partial u^\alpha\, \partial u^\beta} = 0, \qquad 2 \leqq \alpha, \quad \beta \leqq n-1.$$

This condition, together with the fact that

$$\sum_{i=1}^{n-1} u^i \frac{\partial g}{\partial u^i} = g,$$

implies that $S = 0$ if and only if

$$\frac{\partial^2 g}{\partial u^i\, \partial u^j} = 0, \qquad 1 \leqq i, \quad j \leqq n-1.$$

If $S = 0$ then

$$\eta^n = B_3(dx^n - g_1(x)\, dx^1 - \cdots - g_{n-1}(x)\, dx^{n-1})$$

and the condition $d\eta^n \equiv 0 \pmod{\eta^n}$ is precisely the condition that the vector fields $X_i := \partial/\partial x^i + g_i(x)\, \partial/\partial x^n$ all commute. □

REFERENCES

[1] W. M. Boothby and W. P. Dayawansa, *Some global aspects of the feedback linearization problem*, in Differential Geometry: The Interface Between Pure and Applied Mathematics, M. Lucksic, C. Martin, and W. Shadwick, eds., AMS Contemporary Mathematics 68, American Mathematical Society, Providence, RI, 1987, pp. 51-64.

[2] R. Bryant, *On notions of equivalence of variational problems with one independent variable*, in Differential Geometry: The Interface Between Pure and Applied Mathematics, M. Lucksic, C. Martin, and W. Shadwick, eds., AMS Contemporary Mathematics 68, American Mathematical Society, Providence, RI, 1987, pp. 65-76.

[3] E. CARTAN, *Les sous-groupes des groupes continus de transformations*, Oeuvres Complètes II, Ann. Ec. Normale, 25 (1908), pp. 719-856.

[4] ———, *Leçons sur les Invariants Intégraux*, Hermann, Paris, 1922.

[5] R. B. GARDNER, *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. Sussman, eds., Progress in Mathematics 27, Birkhauser, Boston, MA, 1983, pp. 117-180.

[6] ———, *Lectures on the Method of Equivalence with Applications to Control Theory*, CBMS-NSF Regional Conference Series in Applied Mathematics 58, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[7] R. B. GARDNER AND W. F. SHADWICK, *Overdetermined equivalence problems with an application to feedback equivalence*, in Differential Geometry: The Interface between Pure and Applied Mathematics, M. Lucksic, C. Martin, and W. Shadwick, eds., AMS Contemporary Mathematics 68, American Mathematical Society, Providence, RI, 1987, pp. 111-119.

[8] P. GRIFFITHS, *Exterior differential systems and the calculus of variations*, Progress in Mathematics, Birkhauser, Boston, MA, 1983.

[9] R. HERMANN, *The theory of equivalence of Pfaffian systems and input systems under feedback*, Math. Systems Theory, 15 (1982), pp. 343-356.

[10] H. HERMES AND J. P. LASALLE, *Functional analysis and optimal control*, Math. Sci. Engrg., 22, 56 (1969).

[11] A. KRENER, *Normal forms for linear and nonlinear systems*, in Differential Geometry: The Interface Between Pure and Applied Mathematics, M. Lucksic, C. Martin, and W. Shadwick, eds., AMS Contemporary Mathematics 68, American Mathematical Society, Providence, RI, 1987, pp. 157-189.

[12] M. SUTTON, *Equivalence of particle Lagrangians under contact transformations*, Ph.D. dissertation, Mathematics Department, University of North Carolina, Chapel Hill, NC, 1988.

[13] G. WILKENS, *Local feedback equivalence of control systems with three state and two control variables*, Ph.D. dissertation, Mathematics Department, University of North Carolina, Chapel Hill, NC, 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1369-1380, November 1989

008

PARAMETRIZED INTEGRATION OF MULTIFUNCTIONS WITH APPLICATIONS TO CONTROL AND OPTIMIZATION*

ZVI ARTSTEIN†

Abstract. Integration of a set-valued map depending on a parameter is examined. If a point in the range depends measurably on the parameter, then it is the integral of a selection that depends measurably on the parameter. This is proved in the paper and applied in two cases: a control setting, where a Filippov-type lemma for chattering systems is verified; and an optimization problem, where existence of unvarying solutions to asymptotic stochastic maximization is established.

Key words. set-valued maps, integration of multifunctions, selections, chattering controls, Filippov Lemma, optimization

AMS(MOS) subject classifications. 28B20, 93C15, 49A36

1. Introduction. Integration of set-valued functions is a powerful tool in many branches of applied mathematics, including control theory, optimization, statistics, and mathematical economics. Recent investigations in these applications have generated the need for what we term here parametrized integration, namely, the simultaneous integration of multifunctions that depend on a parameter. In this paper we verify a property of the parametrized integration, and present two applications of it.

The property we are interested in is an integral version of the implicit functions lemma used in control theory. It says, roughly, that a point in the range that depends measurably on the parameter can be realized as an integral that depends measurably on the parameter. We state the full result in the next section after recalling the basic definitions and setting the technical framework. The proof of the main result is given in § 5. It uses some known facts about multifunctions, that we recall in § 3, and some lemmas that are proved in § 4.

The two applications are given in §§ 6 and 7. The first application is in control theory; it is the analogue of the Filippov Lemma, here in the context of chattering equations. The second application deals with event depending cost in asymptotic stochastic optimization. We give enough details of these problems to make the arguments of the applications self-contained; however, for the complete background and motivation, the reader is sent elsewhere.

2. The main result. First we recall the notion of integrating set-valued functions, and set the framework in which the main result is stated.

Let S be a measure space, with β a probability measure on it. Let F be a mapping that assigns to each $s \in S$ a subset $F(s)$ of the Euclidean n-dimensional space. An integrable selection (with respect to β) of F is a function f that is β-integrable, and such that $f(s) \in F(s)$ for β-almost every s. The integral of F with respect to β is defined by

$$\int_S F(s)\beta(ds) = \left\{ \int_S f(s)\beta(ds) : f \text{ an integrable selection of } F \right\}. \tag{2.1}$$

This definition was introduced by Aumann [7]. See Castaing and Valadier [11], Hildenbrand [13], Klein and Thompson [14] and references therein for only part of the applicability of this concept.

* Received by the editors May 18, 1988; accepted for publication (in revised form) December 6, 1988.
† Department of Theoretical Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel.

Note that we use $\beta(ds)$ to denote integration with respect to the measure β and the variable s. This is convenient when more parameters are present in the formulas. If no confusion arises, we suppress the dependence on s or on β, and thus write $\int F\,d\beta$ or $\int F$ for the integral.

The framework we are interested in is such that the set-valued function depends on a parameter, say t, in a measure space T with a measure λ on it. Thus, for each $t \in T$ a set-valued map F_t and a probability measure β_t on S are given. For each t the integration of F_t with respect to β_t can be performed, resulting in

$$\Gamma(t) = \int_S F_t(s)\beta_t(ds), \tag{2.2}$$

which is a set in R^n, depending on the parameter t.

The problem we address is the following. Suppose γ is a selection of Γ, namely, $\gamma(t) \in \Gamma(t)$ for λ-almost every t. Then $\gamma(t)$ can be written as $\int f_t(s)\beta_t(ds)$, with $f_t(\cdot)$ a β_t-selection of $F_t(\cdot)$. Then can $f_t(s) = f(t, s)$ be chosen measurable in the two variables? Our answer is stated in the theorem below, after some notations are recalled and the underlying assumptions listed.

Notations. We switch at will between the notation $F_t(s)$ and $F(t, s)$. Multifunction = set-valued function. We write $|x|$ for the Euclidean norm of $x \in R^n$, and $\|K\| = \sup\{|x|: x \in K\}$ for $K \subset R^n$. General definitions for, and general properties of, set-valued functions are stated for a set-valued function G defined on a measure space U with a measure η on it.

The multifunction G is *measurable* if for every closed set C the set $\{u: G(u) \cap C \neq \emptyset\}$ is measurable. This is a standard definition, and it is equivalent to other common definitions in the case that G has closed values, a case that we adopt anyway. See Castaing and Valadier [11, Chap. III]. The set-valued map G is *η-integrably bounded* if there exists an η-integrable scalar function $b: U \to [0, \infty]$ such that $\|G(u)\| \leqq b(u)$ for all $u \in U$.

The space S, on which the probability measures β_t are defined, is assumed to be metric, separable, and complete. We consider the ensemble of probability measures on S as a metric space itself, with the metric being generated by the weak convergence of measures (see, e.g., Billingsley [9], or Hildenbrand [13, p. 48]). Thus, continuity or measurability properties of $t \to \beta_t$ are understood with respect to this metric structure.

The following conditions are assumed throughout.

Underlying Assumptions. (i) S is a complete separable metric space with its Borel structure.

(ii) T is a complete separable metric space, with its Borel structure, and λ is a σ-additive finite measure on T (in particular, λ is tight; see [9, p. 10]).

(iii) The mapping $t \to \beta_t$, which associates with each t a probability measure β_t on S, is measurable.

(iv) The values $F(t, s)$ are nonempty compact subsets of R^n.

(v) $F(\cdot, \cdot)$ is measurable on $T \times S$.

(vi) For each $t \in T$ the mapping $F_t(\cdot)$ is β_t-integrably bounded.

THEOREM. *Let $\gamma: T \to R^n$ be measurable and such that $\gamma(t) \in \Gamma(t)$ for λ-almost every t. Then there exists an $f(t, s)$, measurable on $T \times S$, such that for λ-almost every t, the inclusion $f(t, s) \in F(t, s)$ holds for β_t-almost every s, and $\int f(t, s)\beta_t(ds) = \gamma(t)$.*

3. Preliminary results. For the convenience of the reader here we collect some results about multifunctions and their integrals. We either provide references for a proof or hint how it goes.

Notation. When $x \in R^n$ and $K \subset R^n$ we write $\operatorname{dist}(x, K)$ for $\inf\{|x-y|: y \in K\}$. We denote by $p \cdot x$ the scalar product of p and x. We write $\sup p \cdot K$ for $\sup\{p \cdot x: x \in K\}$, and likewise $\inf p \cdot K = \inf\{p \cdot x: x \in K\}$; we may replace sup by max when the supremum is attained, e.g., when K is compact. The boundary of K in the direction p, namely $\{x \in K: p \cdot x = \sup p \cdot K\}$, is denoted K_p.

In what follows, U is a complete separable metric space with its Borel structure and η is a σ-additive finite measure on U. The multifunctions we treat have subsets of R^n as values.

PROPOSITION 3.1. *Let G be a measurable multifunction with closed nonempty values; then a measurable selection of G exists.*

Proof. See, e.g., Castaing and Valadier [11, p. 65], Hildenbrand [13, p. 54], or Rockafellar [15] for the proof.

PROPOSITION 3.2. *Let G be a measurable multifunction with compact nonempty values. Then $\|G(u)\|$ is measurable. If $h: U \to R^n$ is measurable, then $u \to \operatorname{dist}(h(u), G(u))$ is measurable, and there exists a selection g of G such that $\operatorname{dist}(h(u), G(u)) = |h(u) - g(u)|$.*

Proof. The measurability of $\|G(u)\|$ and $\operatorname{dist}(h(u), G(u))$ follows easily from the Castaing Representation Theorem (see Castaing and Valadier [11, Thm. III.7] or Rockafellar [15, Thm. 1B]). To prove the existence of $g(u)$, consider the multifunction H defined by $H(u) = G(u) \cap \{x: |x - h(u)| = \operatorname{dist}(h(u), G(u))\}$. It is measurable as the intersection of two measurable multifunctions (see Rockafellar [15, Thm. 1M]); any selection of H, existing by the previous result, is the desired g.

PROPOSITION 3.3. *Let G be an integrably bounded measurable multifunction with closed nonempty values. Then $\int G\, d\eta$ is nonempty and compact. If η is atomless then $\int G\, d\eta$ is a convex set.*

Proof. See Aumann [7], or see Klein and Thompson [14, Chap. 18] for the proof.

PROPOSITION 3.4. *Let G_k be a decreasing sequence of multifunctions, namely, $G_k(u) \supset G_{k+1}(u)$ for η-almost every u. Suppose that all G_k are measurable and have closed nonempty values and that G_1 is integrably bounded. Let $G(u) = \bigcap_{k=1}^{\infty} G_k(u)$. Then $\int G\, d\eta = \bigcap_{i=1}^{\infty} \int G_k\, d\eta$.*

Proof. G is measurable, by Rockafellar [15, Thm. 1M], and $\int G$ is trivially included in the intersection of the integrals. To prove the converse let x be a point in the intersection, therefore $x = \int g_k$ with g_k a selection of G_k. By Proposition 3.2 there exist selections h_k of G such that $|g_k(u) - h_k(u)| = \operatorname{dist}(g_k(u), G(u))$. The pointwise limit of the latter is zero, and therefore the Lebesgue Dominated Convergence Theorem implies that $\int (g_k - h_k)$ converge to zero as $k \to \infty$. Namely, $y_k = \int h_k$ converge to x. Since $y_k \in \int G$ and $\int G$ is compact (by Proposition 3.3), it follows that $x \in \int G$.

The following implicit functions property is used several times in the sequel. The above definition of measurability of multifunctions applies also to multifunctions with values being subsets of an abstract metric space.

PROPOSITION 3.5. *Let V be a locally compact separable metric space, and let Q be a multifunction that assigns to each $u \in U$ a closed subset $Q(u)$ of V. Let $\alpha: V \to R^n$ be continuous. Let $G(u) = \{\alpha(v): v \in Q(u)\}$. Suppose Q is measurable and let g be a measurable selection of G. Then there exists a measurable selection q of Q such that $g(u) = \alpha(q(u))$.*

Proof. The result is established under much weaker conditions in Castaing and Valadier [11, Thm. III.38]. Here is a proof in our case. The continuity of α implies easily that the multifunction $H(x) = \{v \in V: \alpha(v) = x\}$ is measurable. Then it is easy to see, e.g., by the Castaing Representation Theorem, that the composition of H and g, namely $Q_1(u) = \{v: \alpha(v) = g(u)\}$, is also a measurable multifunction. Therefore the

intersection $Q(u) \cap Q_1(u)$ generates a measurable multifunction; any selection of it, existing by Proposition 3.1, is a candidate for the desired g. This completes the proof.

In parallel with the definition of the integral of a set-valued function, we write

$$\sum_{i=1}^{\infty} K_i = \left\{ \sum_{i=1}^{\infty} x_i : x_i \in K_i \text{ and the series is summable} \right\}.$$

Note that if K_i is not empty for all i, then $\sum_{i=1}^{\infty} K_i$ is a bounded nonempty set exactly when $\sum_{i=1}^{\infty} \|K_i\| < \infty$. The following corollary is used in the proof of the main result.

COROLLARY 3.6. *For each i let $K_i(u)$ be a measurable multifunction with nonempty compact values in R^n. Suppose that for each u the set*

$$\Gamma_A(u) = \sum_{i=1}^{\infty} K_i(u)$$

is nonempty and bounded. Then Γ_A is a measurable multifunction with compact values. Let $\gamma_A(\cdot)$ be a measurable selection of Γ_A. Then there exist measurable selections $x_i(u)$ of $K_i(u)$ such that $\gamma_A(u) = \sum_{i=1}^{\infty} x_i(u)$.

Proof. Compactness of $\Gamma_A(u)$ is a simple exercise (and a particular case of Proposition 3.3). The measurability of Γ_A follows, e.g., from the Castaing Representation Theorem (see [11, Thm. III.7] or [15, Thm. 1B]). To prove the existence of $x_i(u)$ consider $V = R^n \times R^n$ and define $Q(u) = K_1(u) \times \sum_{i=2}^{\infty} K_i(u)$ and $\alpha(x, y) = x + y$. By Proposition 3.5 there are measurable selections $x_1(u)$ of $K_1(u)$ and $y_2(u)$ of $\sum_{i=2}^{\infty} K_i(u)$ such that $x_1(u) + y_2(u) = \gamma_A(u)$. Inductively, $x_1(u), \cdots, x_{j-1}(u)$ and $y_j(u)$ are defined, and y_j is a measurable selection of $\sum_{i=j}^{\infty} K_j$. Consider $Q(u) = K_j(u) \times \sum_{i=j+1}^{\infty} K_i(u)$. By Proposition 3.5 there are $x_j(u)$ and $y_{j+1}(u)$, selections of $K_j(u)$ and $\sum_{i=j+1}^{\infty} K_i(u)$ such that $x_j(u) + y_{j+1}(u) = y_j(u)$. In this way the sequence of measurable selections $x_1(u)$, $x_2(u) \cdots$ is defined. The boundedness of $K_1(u) + K_2(u) + \cdots$ implies the pointwise convergence of $y_j(u)$ to zero as $j \to \infty$, and hence $\gamma_A(u) = \sum x_i(u)$. This completes the proof.

The following two results are concerned with the boundary of the integral in the direction of a vector $p \in R^n$; see the definitions above.

PROPOSITION 3.7. *Let G be a measurable multifunction with closed values, and let $p \in R^n$. Then*

$$\int (G(u))_p \, d\eta = \left(\int G(u) \, d\eta \right)_p .$$

Furthermore, if g is a selection of G and $p \cdot g(u) < \sup p \cdot G(u)$ on a set of positive η-measure, then $\int g \, d\eta$ does not belong to $(\int G \, d\eta)_p$.

Proof. The proof follows, e.g., from Klein and Thompson [14, Prop. 18.1.8] or Hildenbrand [13, Prop. 6, p. 63].

Recall that the relative boundary of a convex set K is formed by first taking the linear hyperspace (flat) spanned by K, then the boundary of K in this hyperspace.

PROPOSITION 3.8. *Let K be a measurable multifunction with convex compact values. Let γ_N be a measurable selection of K such that $\gamma_N(u)$ is in the relative boundary of $K(u)$ for all u (in particular, $K(u)$ is never a singleton). Then a measurable $p(u) : U \to R^n$ exists such that $p(u) \cdot \gamma_N(u) = \max p(u) \cdot K(u) > \min p(u) \cdot K(u)$, for almost all u.*

Proof. We say that a vector p is parallel to the convex set C in R^n, if p is spanned by vectors of the form $x - y$ with x and y in C. With each convex compact set C and a vector y in the relative boundary of C we associate the set $P(C, y)$ of vectors p such that $|p| = 1$, p is parallel to C, and $p \cdot y = \max p \cdot C$ (namely, the vectors of norm 1, supporting C at y within the span of C). For $i = 1, \cdots, n$ let $\mathcal{C}_i$ be the collection of

compact convex sets of dimensionality i. On $\mathscr{C}_i$ the mapping $(C, y) \to P(C, y)$ has a closed graph, when the family of convex compact sets is endowed with the Hausdorff distance (for the latter see, e.g., [15]). Hence $P(C, y)$ is a measurable multifunction on $C \in \mathscr{C}_i$ and y in the relative boundary of C. But $\mathscr{C}_0, \mathscr{C}_1, \cdots, \mathscr{C}_n$ forms a measurable partition of the space of compact convex sets, e.g., since the dimension of C is a lower semicontinuous function. Hence $P(C, y)$ is a measurable multifunction on its domain. The mapping $u \to K(u)$ was assumed measurable as a set-valued mapping, but this is equivalent to its measurability as a point-valued mapping into the space of compact sets with the Hausdorff distance (see Rockafellar [15, Prop. 1G]). Therefore, the composition $u \to P(K(u), \gamma_N(u))$ is a measurable multifunction. The selection of it guaranteed by Proposition 3.1 is the desired function $p(u)$.

4. Lemmas. Some of the steps toward proving the theorem are collected in this section; some are of interest for their own sake.

LEMMA 4.1. *Let U be a measure space, and let η be a σ-additive measure on it. Let Θ be an ordered set, not necessarily countable. Let G_θ, $\theta \in \Theta$ be a decreasing family of multifunctions, namely, $\theta < \tau$ in Θ implies $G_\theta(u) \subset G_\tau(u)$ for η-almost every u. Suppose that each G_θ is measurable and takes compact subsets of R^n as values. Then there exists a denumerable cofinal sequence G_{θ_i}, namely, for each θ there is an element θ_i in the sequence such that $G_\theta(u) \supset G_{\theta_i}(u)$ for η-almost every u.*

Proof. We can assume that $\|G_\theta\|$ are bounded, say by 1; indeed, it suffices to prove the result for the multifunctions $d(u)G_\theta(u)$ where $d(u) = \min(1, \|G_{\theta_1}(u)\|^{-1})$, with θ_1 fixed. By Proposition 3.2, the function $d(u)$ is measurable.

Let $\{x_i\}$ be a dense sequence in the unit ball of R^n. With each θ in Θ we associate a real number $R(\theta)$ as follows. First, define

$$r_i(\theta, u) = \operatorname{dist}(x_i, G_\theta(u))$$

for $i = 1, 2, \cdots$. These functions are measurable in u (Proposition 3.2), and bounded by 2. Second, define

$$L(\theta, u) = \sum_{i=1}^{\infty} 2^{-i} r_i(\theta, u).$$

Then $L(\theta, u)$ is measurable and bounded by 2. Finally, let

$$R(\theta) = \int_U L(\theta, u)\eta(du).$$

We claim that $R(\theta) > R(\tau)$ implies that $\theta < \tau$ in Θ, and in particular $G_\theta(u) \subset G_\tau(u)$ for η-almost every u. Indeed, the inclusion property implies that the functions $r_i(\theta, u)$ are η-almost everywhere nondecreasing in θ, and hence so are $L(\theta, u)$ and $R(\theta)$. Once the claim is verified, it is clear that the desired cofinal sequence can be determined by a sequence $R(\theta_i)$, which is cofinal with respect to the usual order among the numbers $R(\theta)$, $\theta \in \Theta$; existence of such a sequence is trivial.

We now return to the framework of the main result and the underlying assumptions.

A convenient tool in our derivations is to consider the measure

$$\mu = \int_T \beta_t \, d\lambda, \tag{4.1}$$

which is a measure on $T \times S$ obtained by integrating the probability measures β_t with respect to λ. Namely, if $D \subset T \times S$ is measurable and $D_t = \{s: (t, s) \in D\}$, then

$$\mu(D) = \int_T \beta_t(D_t) \, d\lambda. \tag{4.2}$$

See, e.g., Bourbaki [10], where the integration is developed for locally compact based measures. It is justified in our case since, given $\varepsilon > 0$, by Lusin's Theorem there is a compact $T_\varepsilon \subset T$ with $\lambda(T \setminus T_\varepsilon) < \varepsilon$ and such that $t \to \beta_t$ is continuous on T_ε. By tightness then (see Billingsley [9]) there is a compact $S_\varepsilon \subset S$ such that $\beta_t(S \setminus S_\varepsilon) < \varepsilon$ for all $t \in T_\varepsilon$. Thus on $T_\varepsilon \times S_\varepsilon$ the integration is justified, and as $\varepsilon \to 0$ we get the integration over all T. This reliance on tightness with reference to Bourbaki [10] should be repeated in some of the lemmas below; we leave out the details. (A different approach used to obtain (4.2), as noted by a referee, is to define μ first for simple functions, and rectangulars, and then employ standard completion arguments in measure theory.)

LEMMA 4.2. *A measurable $f: T \times S \to R^n$ satisfies $f(t, s) \in F(t, s)$ for μ-almost every (t, s) if and only if for λ-almost every t, the inclusion $f(t, s) \in F(t, s)$ holds for β_t-almost every s.*

Proof. Applying Bourbaki [10, p. 11] to the function $\rho(t, s) = \text{dist}\,(f(t, s), F(t, s))$, which is measurable by Proposition 3.2, we obtain

$$\int_{T \times S} \rho(t, s)\, d\mu = \int_T d\lambda \int_S \rho(t, s) \beta_t(ds),$$

an equality that is equivalent to the desired conclusion.

Notation. We denote by $\mathcal{B}(S)$ the Borel σ-field on S, and by $\mathcal{B}_\lambda(T)$ the completion of the Borel σ-field on T with respect to the measure λ.

LEMMA 4.3. *For a measurable $E \subset S$ the function $t \to \beta_t(E)$ is $\mathcal{B}_\lambda(T)$-measurable.*

Proof. The indicator function 1_E is measurable; thus $\beta_t(E) = \int_S 1_E \beta_t(ds)$ is $\mathcal{B}_\lambda(T)$ measurable according to Bourbaki [10, p. 11].

LEMMA 4.4. *Define $w: T \times S \to [0, 1]$ by $w(t, s) = \beta_t\{s\}$, namely, assign to an atom s of β_t its weight, and assign the value zero otherwise. Then w is $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$ measurable. In particular, the set*

$$A = \{(t, s): \text{the point } s \text{ is an atom of } \beta_t\} \tag{4.3}$$

is $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$ measurable.

Proof. For each i let $\sigma_i = (E_{i,1}, E_{i,2}, \cdots)$ be a partition of S into measurable subsets, with the diameters of $E_{i,j}$ converging to zero as $i \to \infty$, and nested, namely, a set $E_{i,j}$ is included in one of the elements of σ_k if $k < i$. Define $w_i(t, s) = \beta_t(E_{i,j})$ if $(t, s) \in E_{i,j}$. Then the functions $w_i(\cdot, \cdot)$ are $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$ measurable by the preceding lemma. Since $w(t, s)$ is the pointwise limit of $w_i(t, s)$ as $i \to \infty$, the result is proved.

For each t the measure β_t has only a countable number of atoms. The following results show that these atoms, and the corresponding values of the multifunction F, can be enumerated in a measurable way.

LEMMA 4.5. *There exists a sequence of $\mathcal{B}_\lambda(T)$-measurable functions $q_i(t): T \to S$, each defined possibly on a subset of T, such that there is one-to-one correspondence between the atoms of β_t and the sequence $q_1(t), q_2(t), \cdots$, when defined.*

Proof. Consider the sets

$$A_j = \{(t, s): (j+1)^{-1} < w(t, s) \leqq j^{-1}\}$$

for $j = 0, 1, \cdots$, and where $w(t, s) = \beta_t\{s\}$ is the function introduced in the previous lemma. The sets A_j are $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$ measurable by Lemma 4.4. Each A_i is the graph of a multifunction, say $L_i(t)$, from T into subsets of S, with values consisting of a finite number of points (indeed, $L_i(t)$ cannot have more than $i+1$ elements). Therefore, each L_i is $\mathcal{B}_\lambda(T)$ measurable (see Castaing and Valadier[11, Thm. III.30]). Using the selection theorem, possibly $i+1$ times for L_i, it is easy to obtain $q_{i,1}(t), \cdots, q_{i,i+1}(t)$, functions of the type we want, that exhaust the elements in $L_i(t)$. The ensemble of all these $q_{i,j}$ forms the desired sequence.

COROLLARY 4.6. *There exists a sequence $K_i(t)$ of $\mathcal{B}_\lambda(T)$-measurable multifunctions on T, with values being compact, possibly empty, subsets of R^n, such that for a given t there is a one-to-one correspondence between the nonempty elements $K_1(t)$, $K_2(t), \cdots$, and the values $\beta_t(\{s_i\})F(t, s_i)$ when s_i go over all atoms of β_t. In particular, if we replace the empty values $K_i(t)$ by $\{0\}$ we get*

$$\sum_{i=1}^{\infty} K_i(t) = \int_{A_t} F(t, s)\beta_t(ds) \tag{4.4}$$

with $A_t = \{s$: the point s is an atom of $\beta_t\}$.

Proof. Define $K_i(t) = \beta_t(\{q_i(t)\})F(t, q_i(t))$, with $q_1(t)$, $q_2(t), \cdots$ given in the preceding lemma.

COROLLARY 4.7. *Let $G(t, s)$ be a measurable multifunction on $T \times S$, with compact values in R^n and suppose $G_t(\cdot)$ is β_t-integrably bounded. Then $t \to \int G(t, s)\beta_t(ds)$ is a $\mathcal{B}_\lambda(T)$-measurable multifunction.*

Proof. The integration is the sum $\Gamma_A(t) + \Gamma_N(t)$, with $\Gamma_A(t)$ the result of the integration on the atomic part of β_t, and $\Gamma_N(t)$ the result of the integration on the atomless part of β_t, both justified by Lemma 4.4. The sum would be measurable if each of the components is measurable (see, e.g., Rockafellar [15, Prop. 1J]). The same result, together with Corollary 4.6, imply that $\Gamma_A(t)$ is $\mathcal{B}_\lambda(T)$-measurable. The measurability of $\Gamma_N(t)$ is implied by Bourbaki [10, p. 11], once we recall that $\Gamma_N(t)$ is also the integral of the convex hull of $G(t, s)$, and integration of multifunctions with compact convex values coincides with the Bochner integration into a Banach space (see Klein and Thompson [14, p. 190]).

5. Proof of the theorem. Let $\mu = \int \beta_t \, d\lambda$ be the measure on $T \times S$ as defined in (4.1) and (4.2). Let $\mathcal{F}$ be a family of multifunctions defined on $T \times S$ with the following properties.

(a) Each $G \in \mathcal{F}$ is measurable, with compact values.

(b) $G(t, s) \subset F(t, s)$ for μ-almost every (t, s).

(c) $\int_S G(t, s)\beta_t(ds)$ contains $\gamma(t)$ for λ-almost every t.

The family $\mathcal{F}$ is not empty since $F \in \mathcal{F}$. On $\mathcal{F}$ consider the partial order of inclusion μ-almost everywhere.

CLAIM 1. *There exists a minimal element in $\mathcal{F}$, namely, a $G_0 \in \mathcal{F}$ such that if G_1 is in $\mathcal{F}$ and $G_1(t, s) \subset G_0(t, s)$ for μ-almost every (t, s), then the equality $G_1(t, s) = G_0(t, s)$ holds μ-almost everywhere.*

To verify the claim we use Zorn's Lemma. Let G_θ, $\theta \in \Theta$ be a decreasing family, not necessarily countable, of multifunctions in $\mathcal{F}$. By Lemma 4.1 there exists a denumerable cofinal subsequence G_{θ_i}. Define G by

$$G(t, s) = \bigcap_{i=1}^{\infty} G_{\theta_i}(t, s).$$

By Proposition 3.4, property (c) holds for G; it is clear that properties (a) and (b) hold as well, and thus G is a lower bound for G_θ, $\theta \in \Theta$. By Zorn's Lemma there exists a minimal element. This completes the proof of the claim.

Recall that A (of (4.3)) is the set of atoms of all β_t. Let N be the complement of A in $T \times S$. Both A and N are $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$ measurable by Lemma 4.4.

CLAIM 2. *Let G_0 be a minimal element in $\mathcal{F}$. For μ-almost every (t, s) in N the value $G_0(t, s)$ contains exactly one point, say, $G_0(t, s) = \{g_0(t, s)\}$.*

To prove the claim define

$$\Gamma_A(t) = \int_{A_t} G_0(t, s)\beta_t(ds), \qquad \Gamma_N(t) = \int_{N_t} G_0(t, s)\beta_t(ds) \tag{5.1}$$

and $\Gamma_A(t)=\{0\}$ or $\Gamma_N(t)=\{0\}$ in case the displayed formula is empty (and where A_t and N_t denote the t-sections of A and N). Both integrals are well defined by Lemma 4.4, and both multifunctions are $\mathcal{B}_\lambda(T)$-measurable by Corollary 4.7. Since $G_0 \in \mathcal{F}$ it follows that $\gamma(t) \in \Gamma_A(t)+\Gamma_N(t)$. By Corollary 3.6, for $i=1, 2$, there are selections $\gamma_A(t)$ of $\Gamma_A(t)$ and $\gamma_N(t)$ of $\Gamma_N(t)$ such that

$$\gamma_A(t)+\gamma_N(t)=\gamma(t). \tag{5.2}$$

By Proposition 3.3 the values $\Gamma_N(t)$ are compact and convex. We now examine two cases.

The first case is that for t in a subset, say T_1, of positive measure, the value $\Gamma_N(t)$ is not a singleton, and $\gamma_N(t)$ is in the relative interior of $\Gamma_N(t)$. Suppose, without loss of generality, that the Euclidean distance between $\gamma_N(t)$ and the relative boundary of $\Gamma_N(t)$ is at least $\varepsilon>0$ for $t \in T_1$ (if there is no lower bound for this distance we can consider a subset of T_1 on which there is a lower bound). Let $M \subset (T_1 \times S) \cap N$ be a measurable set with the property that for $(t, s) \in M$ the value $G_0(t, s)$ is not a singleton, and

$$\int_{M_t} \|G_0(t, s)\| \beta_t(ds) < \varepsilon. \tag{5.3}$$

Such a set exists by the β_t-integrability of $\|G_0(t, s)\|$. Define $H(t, s)=G_0(t, s)$ for $(t, s) \notin M$, and $H(t, s)=\{g(t, s)\}$ otherwise, when g is a selection of G_0. Then, in view of (5.3) and the location of $\gamma_N(t)$ in $\Gamma_N(t)$, the integral of H over N_t contains $\gamma_N(t)$, and hence $H \in \mathcal{F}$. But H is strictly smaller than G_0 in the inclusion μ-almost everywhere, a contradiction to the minimality of G_0.

The second case is that for t in a subset of positive measure, say T_2, the value $\Gamma_N(t)$ is not a singleton, and $\gamma_N(t)$ is in the relative boundary of $\Gamma_N(t)$. Let $p(t)$ be given by Proposition 3.8 with $K=\Gamma_N$. For $t \in T_2$ let us define $H(t, s)=G_0(t, s)_{p(t)}$, namely the boundary of $G_0(t, s)$ in the direction $p(t)$, and define $H(t, s)=G_0(t, s)$ if $t \notin t_2$. Then, by Proposition 3.7 we have that $\gamma_N(t)$ belongs to the integral of $H(t, s)$ over N_t with respect to β_t; therefore $H \in \mathcal{F}$. But the integral of G_0 contains more than that of H; hence H is strictly smaller than G_0 on the order of $\mathcal{F}$, which is a contradiction to the minimality of G_0.

Once the two cases are excluded, we conclude that $\Gamma_N(t)$ is a singleton for all t; hence the second claim is verified.

We now proceed with the definition of the desired $f(t, s)$. We define

$$f(t, s)=g_0(t, s) \quad \text{if } (t, s) \in N \tag{5.4}$$

where g_0 is given by Claim 2. We use Lemma 4.5 to determine a sequence $q_1(t)$, $q_2(t), \cdots$ that exhausts the atoms of β_t, and use Corollary 3.6 to determine $x_i(t)$, selections of the values $K_i(t)=\beta_t(q_i(t))G_0(t, q_i(t))$ such that $\Sigma x_i(t)=\gamma_A(t)$, and define

$$f(t, q_i(t))=x_i(t), \qquad i=1, 2, \cdots. \tag{5.5}$$

This defines $f(t, s)$ on $T \times S$ in a measurable way with respect to $\mathcal{B}_\lambda(T) \times \mathcal{B}(S)$. A change of μ-measure zero would make it Borel measurable, and a selection of F, according to Lemma 4.2. The desired equality, namely,

$$\int_S f(t, s)\beta_t(ds)=\gamma(t),$$

follows from (5.4) and (5.5), taking (5.2) into account. This completes the proof.

6. An application in control.

Background. Consider the control system

$$\dot{x} = f(x, t) + g(u, t), \qquad u \in U(t) \tag{6.1}$$

defined on $[t_0, t_1]$. The admissible controls are measurable functions $u(t): [t_0, t_1] \to R^m$ satisfying $u(t) \in U(t)$, the latter being a prescribed multifunction. A reduction can be performed as follows. Consider the system

$$\dot{x} = f(x, t) + v, \qquad v \in V(t) \tag{6.2}$$

where $V(t)$ encompasses the actual effect of the control on the dynamics, namely,

$$V(t) = \{g(u, t): u \in U(t)\}.$$

It is clear that an admissible control of (6.1) generates an admissible control of (6.2), namely, $v(t) = g(u(t), t)$, which yields the same trajectory. The question is whether or not the reduction (6.2) is actually equivalent to (6.1) in the sense that all admissible controls $v(t): [t_0, t_1] \to V(t)$, are generated by measurable functions $u(t): [t_0, t_1] \to U(t)$. The answer is positive, under mild conditions, and this equivalence was observed by Filippov and was the basis for introducing the differential inclusions $\dot{x} \in F(x, t)$ as a convenient model of control systems. See Filippov [12] and Berkovitz [8] for more on the Filippov Lemma and its applications in control theory. The implicit functions property that we quoted in Proposition 3.5 can be used in proving such an equivalence.

A similar equivalence question arises in [3] and [4] in a more complicated situation as follows.

The application. Let S be a separable complete metric space. For each $t \in [t_0, t_1]$ let β_t be a probability measure on S. Let $g(u, t, s): R^m \times [t_0, t_1] \times S \to R^n$ be given. The control system takes the form

$$\dot{x} = f(x, t) + \int_S g(u, t, s)\beta_t(ds), \qquad u \in U(t, s). \tag{6.3}$$

The admissible controls are measurable functions $u(t, s): [t_0, t_1] \times S \to R^m$ satisfying the constraints $u(t, s) \in U(t, s)$, the latter being a prescribed multifunction.

Such strange looking systems arise as variational limits of control systems with highly oscillatory control coefficients. The probability distribution β_t models the occurrence, in the limit, of instantaneous oscillations in the control parameters. This motivates the terminology chattering for such systems. See [3] and [4], for more details.

The system (6.3) can also be reduced to the form

$$\dot{x} = f(x, t) + v, \qquad v \in V(t), \tag{6.4}$$

this time with

$$V(t) = \left\{ \int_S g(u(s), t, s)\beta_t(ds): u \text{ measurable from } S \text{ into } U(t, s) \right\}.$$

And the question is again whether or not each admissible $v(t): [t_0, t_1] \to V(t)$ is generated by a measurable $u(t, s): [t_0, t_1] \times S \to U(t, s)$. We give an answer, employing the theorem of this paper.

Assumptions. Suppose $U(t, s)$ is a measurable multifunction with compact values. Suppose $g(u, t, s)$ is continuous in u and measurable in (t, s). Suppose that for each t the multifunction

$$F(t, s) = \{g(u, t, s): u \in U(t, s)\}$$

is β_t-integrably bounded in the s variable. Suppose $t \to \beta_t$ is measurable, and let λ be the Lebesgue measure on $[t_0, t_1]$.

Application 6.1. Under the assumptions, given an admissible control $v(t)$ of (6.4), there exists an admissible control $u(t, s)$ of (6.3) such that

$$v(t) = \int_S g(u(t, s), t, s)\beta_t(ds)$$

for λ-almost every t in $[t_0, t_1]$.

Proof. It follows from the definition of measurability and the continuity of $g(u, t, s)$ in the u variable that the multifunction $F(t, s)$ is measurable and has compact values. We claim that given a measurable selection $\gamma(t, s)$ of $F(t, s)$, there exists an admissible control $u(t, s)$ such that $\gamma(t, s) = g(u(t, s), t, s)$. The existence of such u follows from the implicit functions property given in Proposition 3.5.

In particular it follows that

$$V(t) = \int_S F(t, s)\beta_t(ds). \tag{6.5}$$

Let $v(t)$ be an admissible control of (6.4), namely, a measurable selection of $V(t)$. By the theorem in this paper (and the assumptions) there exists a selection $\gamma(t, s)$ of $F(t, s)$ such that

$$v(t) = \int_S \gamma(t, s)\beta_t(ds).$$

The previous argument, namely that $\gamma(t, s)$ can be realized as $g(u(t, s), t, s)$ with an admissible control u, completes the proof.

7. An application in optimization. Let T be a probability space, with probability measure λ. For each $t \in T$ and each interger k an optimization problem is given as follows:

$$\begin{aligned} (\mathcal{P}_k) \qquad & \text{maximize} && J(x, t), \\ & \text{subject to} && x \in H(t), \\ & && x = \frac{1}{k}(x_1(t) + \cdots + x_k(t)), \\ & && x_j(t) \in F_j(t) \end{aligned}$$

with constraints $H(t)$ and $F_j(t)$ prescribed, and independent of k. The optimal values and the optimal solutions of $(\mathcal{P}_k)$ may of course depend on the integer k. We are interested in the asymptotic behaviour as $k \to \infty$; in particular we are interested in obtaining an unvarying asymptotic solution, namely, a sequence of decisions

$$x_1(t), x_2(t), \cdots, \qquad x_j(t) \in F_j(t) \tag{7.1}$$

such that for large k the finite sequence $x_1(t), \cdots, x_k(t)$ (whose elements do not depend on k) is a good approximation for a solution. A formal way of expressing what properties an asymptotic solution has follows.

DEFINITION. Denote by $y_k(t)$ the average $(x_1(t) + \cdots + x_k(t))/k$. The sequence in (7.1) is an unvarying asymptotic solution of $(\mathcal{P}_k)$, as $k \to \infty$, if λ-almost everywhere dist $(y_k(t), H(t))$ converge to 0 and if $J(y_k(t), t) - v_k(t)$ converge to zero as $k \to \infty$, where $v_k(t)$ is the optimal value of $(\mathcal{P}_k)$.

Maximization problems like the one we treat, and the need for asymptotic unvarying solutions, arise in stochastic planning and in optimization under uncertainty. In mathematical economics, Arrow and Radner [1] offer a model of this type for decisions in large terms. A survey of relations between probabilistic limit laws for multifunctions and asymptotic values and asymptotic solutions of the problems $(\mathscr{P}_k)$ is presented in [2]. Indeed, existence of unvarying asymptotic solutions depends on asymptotic properties of the constraints $F_j(t)$; these asymptotic properties can be derived from stochastic characteristics of the $F_j(t)$ when interpreted as random variables. This is the connection with stochastic optimization in general, and optimization with stochastic constraints in particular.

In this section we employ the main theorem and derive unvarying asymptotic solutions under conditions weaker than those given in [2]. The framework is as follows.

Let S be a complete separable metric space. Let $F(t, s)$ be a multifunction, with values being compact subsets of R^n. The constraints $F_j(t)$ are obtained as

$$F_j(t) = F(t, s_j(t)) \tag{7.2}$$

with $s_j(t)$ a sequence of measurable functions from T into S; they represent t-dependent samplings of the parameter space S. For each t let β_t be a probability distribution on S. It is assumed in the statement of the application below that the realization $\{s_1(t), \cdots, s_k(t)\}$ is a good sample of β_t in some sense. This is done along the lines of [5] and [6]. In [1] and in [2] it is assumed that $F_j(t) = F(s_j(t))$, with $F(s)$ independent of t, and $s_1(t), s_2(t), \cdots$ are identically distributed and independent. This is a particular case of ours, and implies all the approximation properties that we demand for the sample.

Assumptions. The functional $J(x, t)$ is continuous in x and measurable in t. The multifunction $F(t, s)$ has compact values, is measurable in t and is continuous in s (with respect to the Hausdorff metric), and $F(t, \cdot)$ is bounded. The mapping $t \to \beta_t$ is measurable. For each t the set $\int F(t, s)\beta_t(ds)$ is convex, and has a nonempty intersection with $H(t)$. The multifunction $H(t)$ is measurable, with closed values.

Application 7.1. Denote $\Gamma(t) = H(t) \cap \int F(t, s)\beta_t(ds)$. Then $\Gamma(t)$ has compact values, and by Corollary 4.7, it is $\mathscr{B}_\lambda(T)$-measurable. Let $\Gamma_0(t)$ be the set of $x \in \Gamma(t)$ on which $J(x, t)$ achieves its maximum. The continuity of J in x implies that $\Gamma_0(t)$ is nonempty, and it is easy to verify that Γ_0 is $\mathscr{B}_\lambda(T)$-measurable. Let γ_0 be a selection of Γ_0.

By the theorem, there exists a measurable selection $f(t, s)$ of $F(t, s)$ such that

$$\int f(t, s)\beta_t(ds) = \gamma_0(t) \tag{7.3}$$

for λ-almost every t. Let $\beta_{t,k}$ be the empirical measure on S determined by the sample $\{s_1(t), \cdots, s_k(t)\}$. Suppose that for almost every t the measures $\beta_{t,k}$ converge, in the space of probability measures, to β_t and that $f(t, \cdot)$, when defined on the measure space $(S, \beta_{t,k})$, converge in distribution to $f(t, \cdot)$ when regarded as a function on (S, β_t). Then the sequence

$$x_j(t) = f(t, s_j(t)), \qquad j = 1, 2, \cdots \tag{7.4}$$

forms an unvarying asymptotic solution.

Proof. The weak convergence of $\beta_{t,k}$ to β_t and the continuity of $F(t, s)$ in the s variable imply that $\int \operatorname{co} F(t, s)\beta_{t,k}(ds)$ converge in the Hausdorff metric to $\int \operatorname{co} F(t, s)\beta_t(ds)$, where co K is the convex hull of the set K (see, e.g., Artstein and

Wets [5, Thm. 3.1]). Since $\int F(t, s)\beta_t(ds) = \Gamma(t)$ was assumed convex, and since $\beta_{t,k}$ is determined by the sample $s_1(t), \cdots, s_k(t)$, it follows that

$$\limsup_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} F_j(t) \text{ is included in } \Gamma(t).$$

In other words, $J(\gamma_0(t), t)$ is the maximal value of limits of optimal values $v_k(t)$ of the problems $(\mathcal{P}_k)$. We now show that $J(\gamma_0(t), t)$ is the limit of $J(y_k(t), t)$, where $y_k(t)$ is determined by the decisions $x_j(t)$ of (7.4) (see the definition). From the definition it would then follow that (7.4) forms an unvarying asymptotic solution. In fact, the continuity of $J(\cdot, t)$ implies that it suffices to show that $y_k(t) \to \gamma_0(t)$ for λ-almost every t. But this convergence follows from the assumption of convergence in distribution of $f(t, \cdot)$. Indeed, $y_k(t) = \int f(t, s)\beta_{t,k}(ds)$ while $\gamma_0(t) = \int f(t, s)\beta_t(ds)$. This completes the proof.

Remarks. There is an ad hoc assumption in the preceding application, namely, the assumption of convergence in distribution of the functions $f(t, \cdot)$. It would be desirable to deduce the existence of a selection f with this property from conditions on the data. We can come up with some conditions, e.g., if co $F(t, s)$ is strictly convex and $J(\cdot, t)$ is a convex function; but we do not dwell into these considerations here. Other conditions can be relaxed. The condition that $\Gamma(t)$ is convex can be dropped. Indeed, the sampling process averages out the contributions from the atoms of β_t; a modification of (7.4) along the lines of [2, Scheme 2, p. 74] would produce then an unvarying asymptotic solution. We leave out the details. As mentioned, stochasticity of the samples $s_j(t)$, for instance, if they are independent and identically distributed, would enable us to ease more conditions; for instance, the continuity of $F(t, \cdot)$ is not needed then, and the λ-almost everywhere convergence in distribution of $f(t, \cdot)$, demanded in the application, is then satisfied automatically.

REFERENCES

[1] K. J. ARROW AND R. RADNER, *Allocation of resources in large teams*, Econometrica, 47 (1979), pp. 361-385.

[2] Z. ARTSTEIN, *Limit laws for multifunctions applied to an optimization problem*, in Multifunctions and Integrands, Stochastic Analysis, Approximation and Optimization, G. Salinetti, ed., Lecture Notes in Mathematics 1091, Springer-Verlag, Berlin, 1984, pp. 66-79.

[3] ———, *A variational convergence that yields chattering systems*, Ann. Inst. H. Poincaré, Anal. Non Linéaire, to appear.

[4] ———, *Chattering linear systems: A model for rapidly oscillating coefficients*, Math. Control Signals Systems, to appear.

[5] Z. ARTSTEIN AND R. J-B WETS, *Approximating the integral of a multifunction*, J. Multivariate Anal., 24 (1988), pp. 285-308.

[6] ———, *Decentralized allocation of resources among many producers*, J. Math. Economics, to appear.

[7] R. J. AUMANN, *Integrals of set-valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1-12.

[8] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[9] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[10] N. BOURBAKI, *Elements de Mathematique*, Live VI, Integration, Hermann, Paris, 1959, Chap. 6.

[11] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, New York, 1977.

[12] A. F. FILIPPOV, *On certain questions in the theory of optimal control*, SIAM J. Control, 1 (1962), pp. 76-89.

[13] W. HILDENBRAND, *Core and Equilibria of a Large Economy*, Princeton University Press, Princeton, NJ, 1974.

[14] E. KLEIN AND A. C. THOMPSON, *Theory of Correspondences*, Wiley-Interscience, New York, 1984.

[15] 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 543, Springer-Verlag, Berlin, 1976, pp. 159-207.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1381-1402, November 1989

009

A MATROID-THEORETIC APPROACH TO STRUCTURALLY FIXED MODES OF CONTROL SYSTEMS*

KAZUO MUROTA†

Abstract. The structurally fixed modes for a decentralized control system described in the descriptor form $F\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$, $H\mathbf{y} = C\mathbf{x}$ are investigated under a physically reasonable assumption that the coefficients in the equations are classified into independent physical parameters and dimensionless fixed constants. A necessary and sufficient condition for the existence of structurally fixed modes is given in matroid-theoretic terms; the condition can be tested by an efficient algorithm for the independent-flow problem. The combinatorial canonical form of a layered mixed matrix plays a central role in deriving the condition.

Key words. structurally fixed mode, decentralized control system, layered mixed matrix, combinatorial canonical form, matroid-theoretic algorithm

AMS(MOS) subject classifications. 93, 93A15, 93B55, 15, 05C50

1. Introduction. The concept of fixed modes introduced by Wang and Davison [40] is now recognized as one of the fundamental concepts for the decentralized control, especially with respect to stabilization and pole assignment (see also Anderson and Clements [2], Corfmat and Morse [8], and Davison, Gesing, and Wang [10]). In line with the structural or generic approach to controllability initiated by Lin [22] and developed by Glover and Silverman [14], Kobayashi and Yoshikawa [19], Maeda [25], Shields and Pearson [37], and others, the concept of structurally fixed modes is proposed by Sezer and Šiljak [36] and its combinatorial or graph-theoretic characterizations are given by Sezer and Šiljak [36] and Pichai, Sezer, and Šiljak [33]. See also Reinschke [34].

In this paper, the structurally fixed modes for a decentralized control system described in a descriptor form are investigated in a physically reasonable framework (described in § 3) that has been proposed by Murota [26]-[28], [30] in formulating the structural controllability. That is, the structurally fixed modes are discussed under the assumption that the coefficients in the equations are classified into independent physical parameters and dimensionless fixed constants. A necessary and sufficient condition, of a combinatorial nature, for the existence of structurally fixed modes is derived in § 4 with the aid of the combinatorial canonical form (abbreviated as CCF) of a layered mixed matrix (abbreviated as LM-matrix), which is a mathematical tool useful for systems analysis in general [28], [30] and is described briefly in § 2. Furthermore, it is shown that the derived condition can be tested efficiently by a variant of the matroid-theoretic algorithm for the matroid union/partition problem [12], [21] or for the independent-flow problem [13]; the proposed algorithm, being expressed in terms of an auxiliary graph, is suitable for practical applications in that it is free from the numerical difficulty of rounding errors and is guaranteed to run in polynomial time in the size of the control system in question. The established criterion naturally reduces to the graph-theoretic criterion obtained by Pichai, Sezer, and Šiljak [33] in the case where all the nonzero coefficients can be regarded as independent parameters. It is also mentioned in § 7 that a hierarchical decomposition of a dynamical system with respect to an arbitrarily specified feedback structure is provided by the CCF of an LM-matrix associated with the decentralized system. No explicit reference to matroid

* Received by the editors May 25, 1988; accepted for publication (in revised form) February 1, 1989.
† Department of Mathematical Engineering and Information Physics, University of Tokyo, Tokyo 113, Japan.

theory [41] is made in this paper, although extensive use is made of the results in linear algebra that have been obtained with the aid of matroid theory. In this connection the readers are also referred to [16]–[18], [28] for the use of matroid-theoretic concepts and algorithms in circuit theory.

To be specific, consider a linear time-invariant dynamical system with ν local control stations described as follows:

$$\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{1.1a}$$

$$\mathbf{y}(t) = C\mathbf{x}(t) \tag{1.1b}$$

where $\mathbf{x}$ is the state-vector,

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_\nu \end{pmatrix} \quad \text{and} \quad \mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_\nu \end{pmatrix}$$

are the input-vector and the output-vector, respectively, consisting of the input-vectors $\mathbf{u}_i$ $(i = 1, \cdots, \nu)$ and the output-vectors $\mathbf{y}_i$ $(i = 1, \cdots, \nu)$ of the local control stations. The matrices A, B, and C are real and constant; corresponding to the local stations, B and C are partitioned into ν blocks as

$$B = (B_1 | \cdots | B_\nu), \qquad C = \begin{pmatrix} C_1 \\ \vdots \\ C_\nu \end{pmatrix}.$$

The local output feedback is specified by a block-diagonal real matrix

$$K = \text{block diag}\,[K_1, \cdots, K_\nu], \tag{1.2}$$

where the size of K_i is such that matrix product $B_iK_iC_i$ is defined $(i = 1, \cdots, \nu)$. That is, K represents the nondynamic decentralized output feedback

$$\mathbf{u}(t) = K\mathbf{y}(t), \tag{1.3}$$

i.e., $\mathbf{u}_i = K_i\mathbf{y}_i$ $(i = 1, \cdots, \nu)$. The local output feedback control with dynamic compensation is described by

$$\dot{\mathbf{z}}(t) = L\mathbf{z}(t) + M\mathbf{y}(t), \tag{1.4a}$$

$$\mathbf{u}(t) = N\mathbf{z}(t) + K\mathbf{y}(t) + P\mathbf{v}(t) \tag{1.4b}$$

where

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_\nu \end{pmatrix} \quad \text{and} \quad \mathbf{v} = \begin{pmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_\nu \end{pmatrix},$$

and $\mathbf{z}_i$ and $\mathbf{v}_i$ are, respectively, the state-vector and the external input-vector of the ith feedback controller $(i = 1, \cdots, \nu)$; the matrices L, M, N, and P are block-diagonal real matrices of appropriate sizes.

Let $\mathcal{K}$ be the family of all real matrices K of the form (1.2). The greatest common divisor of the set of characteristic polynomials of $A + BKC$, for all $K \in \mathcal{K}$, is called the *fixed polynomial* of (A, B, C) with respect to $\mathcal{K}$, and denoted by $\psi(s) = \psi(s; A, B, C, \mathcal{K})$, i.e.,

$$\psi(s; A, B, C, \mathcal{K}) = \gcd\{\det(A + BKC - sI) \mid K \in \mathcal{K}\}. \tag{1.5}$$

A complex number $\lambda \in \mathbf{C}$ is called a *fixed mode* of (A, B, C) with respect to $\mathcal{K}$ if λ is an eigenvalue of $A+BKC$ for all $K \in \mathcal{K}$, i.e., if $\psi(\lambda; A, B, C, \mathcal{K}) = 0$. The importance of the concept of fixed modes is demonstrated by the following results due to Wang and Davison [40] and to Corfmat and Morse [8]: (i) the system (1.1) is stabilizable by the decentralized dynamic output feedback (1.4) if and only if all the fixed modes of (A, B, C) have negative real parts; and (ii) the spectrum of the closed-loop system (1.1) and (1.4) is freely assignable by means of $K \in \mathcal{K}$ if and only if there exist no fixed modes for (A, B, C).

As already noted in [40], the notions of fixed polynomial and fixed modes can be defined for (A, B, C) with respect to an arbitrarily specified family $\mathcal{K}$ of the matrices K, not necessarily of the form (1.2). In the special case where $\mathcal{K}$ is composed of all matrices of the compatible size (i.e., of size such that matrix product BKC is defined), the fixed mode with respect to $\mathcal{K}$ is called the *centralized fixed mode* in [10]. It is pointed out in [33] that the above-mentioned result (i) on the stabilizability can be extended to the situation where $\mathcal{K}$ consists of those matrices K that are subject to an arbitrarily specified zero/nonzero structure and where M in (1.4) is constrained to a structure consistent with that of K.

A fixed mode of (A, B, C) is called a *structurally fixed mode* if it stems not from accidental matching of the numerical values of system parameters but from the structure of the system [36]. In [33] and [36], however, the structurally fixed modes are defined with respect to the zero/nonzero structure of A, B, and C. Namely, the fixed modes are considered for the family $\mathcal{S}$ of systems that are "structurally equivalent" to (A, B, C), where a system $(\hat{A}, \hat{B}, \hat{C})$ is called structurally equivalent to (A, B, C) if $\hat{A}$, $\hat{B}$, and $\hat{C}$ have, respectively, the same zero/nonzero structure as that of A, B, and C. Obviously, the concept of structurally fixed modes can be defined with respect to a pair $(\mathcal{S}, \mathcal{K})$ of a more general family $\mathcal{S}$ of systems and a more general family $\mathcal{K}$ of output feedbacks.

It has been gradually recognized that in the design and analysis of large-scale systems in general, "structural" or qualitative considerations are no less important than "numerical" or quantitative ones. It should be emphasized that the "structure" should be defined on a physically sound basis, and that it is crucial in this respect to work with elementary mathematical representations of physical structures expressed in terms of elementary variables, and not with sophisticated and compact representations.

In the present case of dynamical systems, therefore, the "standard state-space form" (1.1) is not suitable for expressing the elementary physical structure in that the entries of the matrices A, B, and C usually have mutual algebraic relations among themselves. The so-called "descriptor form" (cf. Armentano [4], Cobb [6], [7], Luenberger [23], [24], Verghese, Lévy, and Kailath [38]) is more suitable. That is, (1.1) and (1.3) are to be replaced, respectively, by

$$F\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{1.6a}$$

$$H\mathbf{y}(t) = C\mathbf{x}(t), \tag{1.6b}$$

$$G\mathbf{u}(t) = K\mathbf{y}(t). \tag{1.7}$$

Then, as advocated by Murota [28] and Murota and Iri [31], it is often justified to assume that the coefficients in these equations are classified into two groups, one of independent physical parameters and the other of fixed constants, typically simple integers such as ± 1.

In this paper we recognize the structure of a system in the above sense. That is, we set $\mathscr{S}$ to be the family of systems (A, B, C, F, H) of (1.6), in which the independent parameters of those coefficient matrices can take arbitrary real values while the fixed coefficients remain constants. As for the feedback structure constraint we assume that $\mathscr{K}$ consists of pairs (G, K) of the matrices G and K having structures specified similarly in terms of independent parameters and fixed constants. We will discuss the structurally fixed modes with respect to such $(\mathscr{S}, \mathscr{K})$ (see § 3 for the precise formulation of the problem). Combinatorial and algorithmic characterizations of the existence of structurally fixed modes will be established in Theorem 4.2 and Theorem 5.1.

The present formulation of the fixed modes for a general descriptor system (1.6) includes those treated in [8], [10], [33], [36], and [40] as special cases. When a decentralized control system is described by (1.6) with F and H being nonsingular, and $\mathscr{K}$ consists of all block-diagonal matrices (G, K) compatible with the locality of admissible feedbacks, the fundamental results of Wang and Davison [40] and Corfmat and Morse [8] mentioned above remain valid and clarify the control-theoretic significance of the fixed modes in the present formulation. In this case, nothing new is introduced to the concept of fixed modes of a particular system. Nevertheless, the descriptor form is more suitable for considering structurally fixed modes since it represents the physical structure more faithfully. A general descriptor system, however, may have the so-called impulsive modes that bring about complications [4], [7], [38]. Although substantial results seem to have been obtained so far on the pole assignment by state-feedback [3], [5], [6], the present author is not informed of results on pole assignment by dynamic compensation for a general descriptor system.

2. Preliminaries on mixed matrices. Let $\mathbf{K} \subseteq \mathbf{F}$ be fields. A matrix A is called a *mixed matrix* with respect to $\mathbf{K}$ if

$$A = Q + T \tag{2.1}$$

where (i) $Q = (Q_{ij})$ is a matrix over $\mathbf{K}$, and (ii) $T = (T_{ij})$ is a matrix over $\mathbf{F}$ such that the set $\mathscr{N}(T)$ is (collectively) algebraically independent [39] over $\mathbf{K}$, where $\mathscr{N}(\cdot)$ denotes in general the set of nonzero entries of a matrix. The following identity is fundamental where, for a matrix M in general, we denote the row-set and the column-set of M by Row (M) and Col (M), respectively, and the submatrix with row-set I and column-set J by $M[I, J]$; note also term-rank of M coincides with its generic rank if $\mathscr{N}(M)$ is algebraically independent. See Murota [28], [29], Murota and Iri [31], and Murota, Iri, and Nakamura [32] for the details of this section.

LEMMA 2.1. *For a mixed matrix* $A = Q + T$,

$$\text{rank } A = \max \{\text{rank } Q[R - I, C - J] + \text{term-rank } T[I, J] \mid I \subseteq R, J \subseteq C\},$$

where $R = \text{Row } (A)$, $C = \text{Col } (A)$.

A matrix A is called a *layered mixed matrix* (or an LM-matrix) with respect to $\mathbf{K}$ if it takes the following form (possibly after a permutation of rows):

$$A = \begin{pmatrix} Q \\ T \end{pmatrix} \tag{2.2}$$

and Q and T of (2.2) meet the requirements (i) and (ii) above.

By the *admissible transformation* for an LM-matrix A of (2.2) we mean the transformation of the form:

$$P_r \begin{pmatrix} S & O \\ O & I \end{pmatrix} \begin{pmatrix} Q \\ T \end{pmatrix} P_c \tag{2.3}$$

where S is a nonsingular matrix over the subfield $\mathbf{K}$, and P_r and P_c are permutation matrices. The admissible transformation brings an LM-matrix into another LM-matrix and two LM-matrices are said to be LM-*equivalent* if and only if they are connected by an admissible transformation.

There exists a finest block-triangular matrix, called the *combinatorial canonical form* (or CCF for short), among the matrices that are LM-equivalent to each other. Note that the CCF is a generalization of the canonical decomposition due to Dulmage and Mendelsohn [11] of a bipartite graph, or of a formal incidence matrix [35]. The CCF for a nonsingular LM-matrix is described as follows.

Let $\bar{A}$ be the CCF of a nonsingular A; Row $(\bar{A})$ and Col $(\bar{A})$ are, respectively, partitioned into nonempty blocks as

$$\{R_1, \cdots, R_r\}, \qquad \{C_1, \cdots, C_r\} \tag{2.4}$$

where

$$\emptyset \neq R_k \subseteq \text{Row}\,(\bar{A}), \quad \emptyset \neq C_k \subseteq \text{Col}\,(\bar{A}) \quad \text{for } k = 1, \cdots, r,$$
$$R_k \cap R_l = \emptyset, \quad C_k \cap C_l = \emptyset \quad \text{for } k \neq l.$$

LEMMA 2.2. *The* CCF $\bar{A}$ *of a nonsingular* LM-*matrix* A *has the following properties*:
(1) $\bar{A}$ *is block-triangularized with respect to the partitions* (2.4). *That is,*

$$\bar{A}[R_k, C_l] = O \quad \text{if } 1 \leqq l < k \leqq r.$$

(2) $$\text{rank}\,\bar{A}[R_k, C_k] = |R_k| = |C_k| > 0 \quad \text{for } k = 1, \cdots, r.$$

(3) $\bar{A}$ *is the finest block-triangular matrix with property* (2) *that is* LM-*equivalent to* A.

A nonsingular LM-matrix A will be called (LM)-*irreducible* if its CCF does not split into more than one block, that is, if $r = 1$ above. Each diagonal block $\bar{A}[R_k, C_k]$ of the CCF above is irreducible $(k = 1, \cdots, r)$. The following result of Murota [29] plays the central role in § 4. It is a generalization of the result of Ryser [35] for formal incidence matrices.

LEMMA 2.3. *Let* $A = \begin{pmatrix} Q \\ T \end{pmatrix}$ *be a nonsingular irreducible* LM-*matrix with respect to* $\mathbf{K}$, *and* $\mathcal{T} = \mathcal{N}(T)$ *denote the set of nonzero entries of* T.
(1) det A *is an irreducible polynomial in the ring* $\mathbf{K}[\mathcal{T}]$.
(2) *Each element of* $\mathcal{T}$ *appears in* det A.

3. Problem formulation. For the descriptor system (1.6) and (1.7) we define the fixed polynomial $\psi(s)$ with respect to a family $\mathcal{K}$ of allowable feedbacks (G, K) by

$$\psi(s) = \psi(s; A, B, C, F, H, \mathcal{K}) = \gcd\{\det D(s) \,|\, (G, K) \in \mathcal{K}\} \tag{3.1}$$

where

$$D = D(s) = \begin{pmatrix} A - sF & B & O \\ O & -G & K \\ C & O & -H \end{pmatrix} \tag{3.2}$$

and gcd is considered in $\mathbf{C}[s]$. Note that this is an extension of (1.5) since

$$\det(A + BKC - sI) = \det \begin{pmatrix} A - sI & B & O \\ O & -I & K \\ C & O & -I \end{pmatrix}. \tag{3.3}$$

We call a complex number $\lambda \in \mathbf{C}$ a fixed mode of (A, B, C, F, H) with respect to $\mathcal{K}$ if $\psi(\lambda) = 0$.

We now introduce the first physical observation that explains how we recognize the structure of a system. This will constitute the basis of our choice of the family $\mathcal{S}$ of systems. When a dynamical system is written in the form (1.6) and (1.7) in terms of elementary physical variables, it is often justified to assume that the nonzero entries of the coefficient matrices A, B, C, F, and H are classified into two groups, one of generic parameters and the other of fixed constants. In other words, as observed by Murota and Iri [31], we can distinguish two kinds of numbers that characterize physical systems as follows: (i) those numbers representing independent physical parameters such as resistances in electrical networks that, being contaminated by various noises and errors, take inaccurate values independent of one another, so that they can be modeled as algebraically independent generic numbers, and (ii) those numbers accounting for various sorts of conservation laws such as Kirchhoff's, that, stemming from topological incidence relations, are accurate (often ± 1) in value so that no serious numerical difficulty arises in arithmetic operations on them. See [31] or Chapter 4 of [28] for further discussions.

Based on this physical observation we assume that coefficient matrices A, B, C, F, and H are expressed as follows:

$$\begin{aligned} A = Q_A + T_A, \quad B = Q_B + T_B, \quad C = Q_C + T_C, \\ F = Q_F + T_F, \qquad H = Q_H + T_H, \end{aligned} \tag{3.4}$$

where Q_A, Q_B, etc., are matrices over $\mathbf{Q}$ (the field of rational numbers), and the sets $\mathcal{N}(T_A)$, $\mathcal{N}(T_B)$, etc., of nonzero entries are disjoint and

$$\mathcal{S} = \mathcal{N}(T_A) \cup \mathcal{N}(T_B) \cup \mathcal{N}(T_C) \cup \mathcal{N}(T_F) \cup \mathcal{N}(T_H) (\subseteq \mathbf{R}) \tag{3.5}$$

is algebraically independent over $\mathbf{Q}$. It should be clear that assuming algebraic independence of $\mathcal{S}$ is equivalent to regarding the members of $\mathcal{S}$ as independent parameters, and therefore to considering the family, to be denoted also by $\mathcal{S}$, of systems parametrized by those parameters in $\mathcal{S}$. Such a particular system has a fixed mode with respect to $\mathcal{K}$ if and only if each member of $\mathcal{S}$ has a fixed mode with respect to $\mathcal{K}$.

The feedback structure constraint $\mathcal{K}$ is assumed to be specified by means of a pair of mixed matrices

$$G = Q_G + T_G, \qquad K = Q_K + T_K. \tag{3.6}$$

That is, $\mathcal{K}$ is composed of (G, K) of (3.6), where the nonzero entries of T_G and T_K take any real values.

In this paper we are primarily concerned with combinatorial and algorithmic characterizations of the condition

$$\psi(s; A, B, C, F, H, \mathcal{K}) \in \mathbf{C} \tag{3.7}$$

for a system (A, B, C, F, H) such that

(A1) $\mathcal{S}$ of (3.5) is algebraically independent over $\mathbf{Q}$.

Remark 3.1. The rationality of the entries of Q_A, Q_B, etc., is not essential to the subsequent arguments. In case nonrational constants are involved, we may choose as $\mathbf{K}$ an appropriate extension field of $\mathbf{Q}$. The subfield $\mathbf{K}$ affects the computational complexity of the algorithm to be described in § 5.

Since the members of $\mathcal{N}(T_G) \cup \mathcal{N}(T_K)$ are independent parameters, (A1) implies that D of (3.2) is a mixed matrix (cf. § 2) with respect to $\mathbf{K} = \mathbf{Q}(s)$, i.e.,

$$D = Q_D + T_D \tag{3.8}$$

with

$$Q_D = Q_D(s) = \begin{pmatrix} Q_A - sQ_F & Q_B & O \\ O & -Q_G & Q_K \\ Q_C & O & -Q_H \end{pmatrix}, \tag{3.9}$$

$$T_D = T_D(s) = \begin{pmatrix} T_A - sT_F & T_B & O \\ O & -T_G & T_K \\ T_C & O & -T_H \end{pmatrix}. \tag{3.10}$$

The second physical observation made by Murota [26] (see also [27], [28, Chap. 4]) is that the fixed constants, or the accurate numbers, usually represent topological and/or geometrical incidence coefficients that have no physical dimensions. Therefore, it is natural to expect that the entries of Q_A, Q_B, etc., are dimensionless constants. On the other hand, the indeterminate s should have the physical dimension of the inverse of time, since it corresponds to the differentiation with respect to time. When combined with the principle of dimensional homogeneity [15], [20], this implies that

$$Q_D(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_d}] \cdot Q_D(1) \cdot \mathrm{diag}\,[s^{-c_1}, \cdots, s^{-c_d}] \tag{3.11}$$

for some integers r_i and c_i $(i = 1, \cdots, d)$, where d is the size of the matrix D (see [26] and [28] for details). Note that $-r_i$ and $-c_i$ admit the natural physical interpretation of the exponents to the dimension of time associated, respectively, with the ith row (equation) and the ith column (variable) of D. It is known that (3.11) holds for some integers r_i and c_i $(i = 1, \cdots, d)$ if and only if

(A2) Every nonvanishing subdeterminant of $Q_D(s)$ is a monomial in s over $\mathbf{Q}$ (i.e., of the form αs^p with a rational number α and an integer p).

Our problem is to derive necessary and sufficient conditions for (3.7) under the assumptions (A1) and (A2).

4. Algebraic/combinatorial characterization of a fixed polynomial. We will solve our problem formulated in § 3 in a still more general form as follows. (Examples in § 6 will provide a concrete idea to the argument below.) Let s be an indeterminate over $\mathbf{C}$ and let

$$D = D(s) = Q_D + T_D \tag{4.1}$$

be a $d \times d$ nonsingular matrix such that

$$Q_D = Q_D(s) = Q^0 + sQ^1, \tag{4.2}$$

$$T_D = T_D(s) = (T^0 + sT^1) + \hat{K} \tag{4.3}$$

where

(A1) $\mathcal{S} \equiv \mathcal{N}(T^0) \cup \mathcal{N}(T^1)$ (disjoint union) $(\subseteq \mathbf{C})$ is algebraically independent over $\mathbf{Q}$.

(A2) Every nonvanishing subdeterminant of $Q_D(s)$ is a monomial in s over $\mathbf{Q}$.

(A3) The elements of $\mathcal{K} \equiv \mathcal{N}(\hat{K})$ are indeterminates over $\mathbf{C}(s)$.

Recall that (A2) implies

$$Q_D(s) = \mathrm{diag}\,[s^{r_1}, \cdots, s^{r_d}] \cdot Q_D(1) \cdot \mathrm{diag}\,[s^{-c_1}, \cdots, s^{-c_d}] \tag{4.4}$$

for some integers r_i and c_i $(i = 1, \cdots, d)$.

Remark 4.1. It should be clear that for the problem formulated in § 3 we have

$$Q^0=\begin{pmatrix} Q_A & Q_B & O \\ O & -Q_G & Q_K \\ Q_C & O & -Q_H \end{pmatrix}, \qquad Q^1=\begin{pmatrix} -Q_F & O & O \\ O & O & O \\ O & O & O \end{pmatrix},$$

$$T^0=\begin{pmatrix} T_A & T_B & O \\ O & O & O \\ T_C & O & -T_H \end{pmatrix}, \qquad T^1=\begin{pmatrix} -T_F & O & O \\ O & O & O \\ O & O & O \end{pmatrix},$$

$$\hat{K}=\begin{pmatrix} O & O & O \\ O & -T_G & T_K \\ O & O & O \end{pmatrix}$$

for the matrices introduced here.

In accordance with (3.1) we define the fixed polynomial $\psi(s)$ as the greatest common divisor in $\mathbf{C}[s]$ of all $\det D(s)$, where arbitrary real values are substituted into $\mathcal{K}$. Also we call a complex number $\lambda \in \mathbf{C}$ a fixed mode with respect to $\mathcal{K}$ if $\psi(\lambda)=0$.

Regarding $\det D(s)$ as a polynomial in $(s, \mathcal{S}, \mathcal{K})$ over $\mathbf{Q}$, let

$$\det D(s)=\prod_{k\in\Psi_0}\psi_k(s)\cdot\prod_{k\in\Psi_1}\psi_k(s,\mathcal{S})\cdot\prod_{k\in\Psi_2}\psi_k(s,\mathcal{S},\mathcal{K}) \tag{4.5}$$

be the decomposition into irreducible polynomials in $\mathbf{Q}[s, \mathcal{S}, \mathcal{K}]$, where $\psi_k(s)\in\mathbf{Q}[s]$ for $k\in\Psi_0$, $\psi_k(s,\mathcal{S})\in\mathbf{Q}[s,\mathcal{S}]-\mathbf{Q}[s]$ for $k\in\Psi_1$, and $\psi_k(s,\mathcal{S},\mathcal{K})\in\mathbf{Q}[s,\mathcal{S},\mathcal{K}]-\mathbf{Q}[s,\mathcal{S}]$ for $k\in\Psi_2$. The following would be obvious.

LEMMA 4.1. *The fixed polynomial* $\psi(s)\in\mathbf{C}[s]$ *is given as*

$$\psi(s)=\prod_{k\in\Psi_0}\psi_k(s)\cdot\prod_{k\in\Psi_1}\psi_k(s,\mathcal{S}). \tag{4.6}$$

Proof. It suffices to show that $\psi_k(\lambda,\mathcal{S},\mathcal{K})$ is not equal to zero as a polynomial in $\mathcal{K}$ over $\mathbf{C}$ for all $\lambda\in\mathbf{C}$ and for all $k\in\Psi_2$. Suppose, to the contrary, there exist $\lambda\in\mathbf{C}$ and $k\in\Psi_2$ such that $\psi_k(\lambda,\mathcal{S},\mathcal{K})=0$ in $\mathbf{C}[\mathcal{K}]$. As a polynomial in $(s,\mathcal{K})$ over $\mathbf{F}\equiv\mathbf{Q}(\mathcal{S})$, ψ_k is expressed as

$$\psi_k(s,\mathcal{S},\mathcal{K})=\sum_{\alpha}\rho_\alpha(s)\mathcal{K}^\alpha$$

where α denotes a multi-index, $\mathcal{K}^\alpha$ is a product of some elements in $\mathcal{K}$, and $\rho_\alpha(s)\in\mathbf{F}[s]$. Note that there exist more than one term in this expression since $\psi_k(s,\mathcal{S},\mathcal{K})\in\mathbf{Q}[s,\mathcal{S},\mathcal{K}]-\mathbf{Q}[s,\mathcal{S}]$ is irreducible. Then $\rho_\alpha(\lambda)=0$ for all α. This means [39] that $\rho_\alpha(s)$, for each α, is divisible in $\mathbf{F}[s]$ by the minimal polynomial of λ over $\mathbf{F}$. However, this contradicts the assumption since the irreducibility of $\psi_k(s,\mathcal{S},\mathcal{K})$ in $\mathbf{Q}[s,\mathcal{S},\mathcal{K}]$ implies [39] its irreducibility as a polynomial in $(s,\mathcal{K})$ over $\mathbf{F}=\mathbf{Q}(\mathcal{S})$.

Consider the first factor on the right-hand side of (4.5). It follows from (A2), or (4.4), that

$$\prod_{k\in\Psi_0}\psi_k(s)=\alpha s^p \qquad (\alpha\in\mathbf{Q}, p\in\mathbf{Z}). \tag{4.7}$$

LEMMA 4.2. $D(s)$ *is nonsingular* (s: *indeterminate*) *with* $p=0$ *in* (4.7) *if and only if* $D(0)$ *is nonsingular.*

The nonsingularity of $D(0)$ can be characterized combinatorially by Lemma 2.1 since $D(0)=Q^0+(T^0+\hat{K})$ is a mixed matrix with respect to $\mathbf{Q}$. (More generally it is

possible to compute the value of p by a matroid-theoretic algorithm. See Theorem 4.2 below.)

The remaining factors in (4.5) are determined as follows. Put

$$\tilde{D} = \tilde{D}(s; t_1, \cdots, t_d) = \begin{pmatrix} I_d & Q_D \\ -\mathrm{diag}\,[t_1, \cdots, t_d] & T_D \end{pmatrix} = \begin{pmatrix} \tilde{Q} \\ \tilde{T} \end{pmatrix} \tag{4.8}$$

where I_d is the unit matrix of order d, and t_i $(i = 1, \cdots, d)$ are indeterminates (distinct from s and $\mathcal{K}$). Obviously,

$$\det D(s) = \det \tilde{D}(s; 1, \cdots, 1). \tag{4.9}$$

The matrix $\tilde{D}$ is an LM-matrix with respect to $\mathbf{K} = \mathbf{Q}(s)$. Let

$$\bar{D} = \bar{D}(s; t_1, \cdots, t_d) = \begin{pmatrix} \bar{D}_1 & \bar{D}_{12} & \cdots & \bar{D}_{1,r-1} & \bar{D}_{1r} \\ O & \bar{D}_2 & \cdots & \bar{D}_{2,r-1} & \bar{D}_{2r} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & \bar{D}_{r-1} & \bar{D}_{r-1,r} \\ O & O & \cdots & O & \bar{D}_r \end{pmatrix} \tag{4.10}$$

be the CCF of $\tilde{D}$ (cf. § 2). Since $\bar{D}$ is obtained from $\tilde{D}$ by the transformation (2.3) with nonsingular S over $\mathbf{Q}(s)$, we have $\det \bar{D} = \rho_0(s) \cdot \det \tilde{D}$ with $\rho_0(s) \in \mathbf{Q}(s) - \{0\}$. Furthermore, we may assume, by virtue of (A2), that $\rho_0(s)$ is a (possibly negative) power of s, i.e.,

$$\begin{aligned} \det \bar{D}(s; t_1, \cdots, t_d) &= \prod_{k=1}^{r} \det \bar{D}_k(s; t_1, \cdots, t_d) \\ &= \alpha_0 s^{p_0} \cdot \det \tilde{D}(s; t_1, \cdots, t_d) \quad (\alpha_0 \in \mathbf{Q}(s) - \{0\}, p_0 \in \mathbf{Z}) \end{aligned} \tag{4.11}$$

and that each entry of $\bar{D}$ belongs to $\mathbf{Q}[s, \mathcal{S}, \mathcal{K}, t_1, \cdots, t_d]$.

LEMMA 4.3. *For* $k = 1, \cdots, r$,

$$\det \bar{D}_k(s; 1, \cdots, 1) = \rho_k(s) \cdot \bar{\psi}_k(s) \tag{4.12}$$

where $\bar{\psi}_k(s) \in \mathbf{Q}[s, \mathcal{S}, \mathcal{K}] - \mathbf{Q}[s]$ *is irreducible and* $\rho_k(s) \in \mathbf{Q}[s]$.

Proof. Denote by $\mathcal{T}_i$ the set of elements of $\mathcal{T} \equiv \mathcal{S} \cup \mathcal{K}$ contained in the row of $\tilde{D}$ corresponding to t_i $(i = 1, \cdots, d)$ (cf. (4.8)). Put $f(\mathcal{T}_1, \cdots, \mathcal{T}_d; t_1, \cdots, t_d) = \det \bar{D}_k$ that is irreducible in $\mathbf{Q}(s)[\mathcal{T}_1, \cdots, \mathcal{T}_d, t_1, \cdots, t_d]$ by Lemma 2.3(1). From the expression

$$f(\mathcal{T}_1, \cdots, \mathcal{T}_d; t_1, \cdots, t_d) = \left(\prod_{t_i \in \mathcal{N}(\bar{D}_k)} t_i \right) \cdot f\left(\frac{\mathcal{T}_1}{t_1}, \cdots, \frac{\mathcal{T}_d}{t_d}; 1, \cdots, 1 \right)$$

(where the notation $\mathcal{T}_i / t_i$ $(i = 1, \cdots, d)$ on the right-hand side means substituting a / t_i for each indeterminate $a \in \mathcal{T}_i$), it follows that $\det \bar{D}_k(s; 1, \cdots, 1)$ is irreducible in $\mathbf{Q}(s)[\mathcal{S}, \mathcal{K}]$, completing the proof. □

When index sets $\bar{\Psi}_1$ and $\bar{\Psi}_2$ are defined by

$$\bar{\Psi}_1 = \{k \mid \bar{D}_k \text{ contains a variable of } \mathcal{S} \text{ and no variable of } \mathcal{K}\}, \tag{4.13a}$$

$$\bar{\Psi}_2 = \{k \mid \bar{D}_k \text{ contains a variable of } \mathcal{K}\}, \tag{4.13b}$$

Lemma 2.3 implies that

$$\bar{\psi}_k(s) = \bar{\psi}_k(s, \mathcal{S}) \in \mathbf{Q}[s, \mathcal{S}] - \mathbf{Q}[s], \qquad k \in \bar{\Psi}_1, \tag{4.14a}$$

$$\bar{\psi}_k(s) = \bar{\psi}_k(s, \mathcal{S}, \mathcal{K}) \in \mathbf{Q}[s, \mathcal{S}, \mathcal{K}] - \mathbf{Q}[s, \mathcal{S}], \qquad k \in \bar{\Psi}_2 \tag{4.14b}$$

where $\bar{\psi}_k(s)$ is defined by (4.12). Combining (4.5), (4.9), (4.11), (4.12), and (4.14), we see that

$$\{\psi_k(s, \mathcal{S}) \mid k \in \Psi_1\} = \{\bar{\psi}_k(s, \mathcal{S}) \mid k \in \bar{\Psi}_1\}, \tag{4.15a}$$

$$\{\psi_k(s, \mathcal{S}, \mathcal{K}) \mid k \in \Psi_2\} = \{\bar{\psi}_k(s, \mathcal{S}, \mathcal{K}) \mid k \in \bar{\Psi}_2\} \tag{4.15b}$$

where in these expressions two polynomials are considered equal if they are equal up to a nonzero multiplicative factor in $\mathbf{Q}$. To sum up, the decomposition (4.5) of $\det D(s)$ into irreducible factors is determined, up to a factor of monomial of s, by the CCF of $\tilde{D}$ of (4.8).

THEOREM 4.1. *Assume* (A1), (A2), *and* (A3) *as well as the nonsingularity of* $D(s)$ *of* (4.1). *The fixed polynomial* $\psi(s)$ *is given by*

$$\psi(s) = \alpha_1 s^{p_1} \prod_{k \in \bar{\Psi}_1} \bar{\psi}_k(s, \mathcal{S}) = \alpha_2 s^{p_2} \prod_{k \in \bar{\Psi}_1} \det \bar{D}_k(s),$$

where $\alpha_i \in \mathbf{Q} - \{0\}$ *and* $p_i \in \mathbf{Z}$ *for* $i = 1, 2$.

Proof. This follows from Lemma 4.1 and (4.15). □

This theorem shows that there exist no nonzero fixed modes if and only if $\bar{\psi}_k(s, \mathcal{S}) \in \mathbf{Q}[\mathcal{S}]$ for all $k \in \bar{\Psi}_1$. This condition can be stated in combinatorial terms as follows.

Recalling that $\bar{D}_k$ is an LM-matrix with respect to $\mathbf{Q}(s)$, express it in the form of (2.2) as

$$\bar{D}_k = \begin{pmatrix} \bar{Q}_k \\ \bar{T}_k \end{pmatrix}. \tag{4.16}$$

Put $\bar{C} = \mathrm{Row}\,(D) \cup \mathrm{Col}\,(D)$ and define $\zeta : \bar{C} \to \mathbf{Z}$ with reference to (4.4) by

$$\zeta(j) = \begin{cases} -r_j & \text{if } j \in \mathrm{Row}\,(D), \\ -c_j & \text{if } j \in \mathrm{Col}\,(D), \end{cases} \tag{4.17}$$

and

$$\zeta(J) = \sum_{j \in J} \zeta(j), \qquad J \subseteq \bar{C}.$$

Then

$$\deg_s \det \bar{Q}_k[\mathrm{Row}\,(\bar{Q}_k), J] = p_0 + \zeta(J) \tag{4.18}$$

for all $J \subseteq \bar{C}_k \equiv \mathrm{Col}\,(\bar{D}_k) \subseteq \bar{C}$ such that $\bar{Q}_k[\mathrm{Row}\,(\bar{Q}_k), J]$ is nonsingular, where p_0 is independent of J.

For $J \subseteq \bar{C}_k$ such that $\bar{T}_k[\mathrm{Row}\,(\bar{T}_k), J]$ is nonsingular, we denote by $\xi_k(J)$ and $\eta_k(J)$ the highest and lowest degrees in s of a nonzero term in $\det \bar{T}_k[\mathrm{Row}\,(\bar{T}_k), J]$.

(Note $\xi_k(J)$ and $\eta_k(J)$ can be computed by solving a bipartite weighted-matching problem. See also § 5.)

Then, we have

$$\deg_s \bar{\psi}_k(s) = \max\{\zeta(\bar{C}_k - J) + \xi_k(J) \mid J \in \mathcal{B}_k\} - \min\{\zeta(\bar{C}_k - J) + \eta_k(J) \mid J \in \mathcal{B}_k\} \tag{4.19}$$

where

$$\mathcal{B}_k = \{J \subseteq \bar{C}_k \mid \bar{Q}_k[\mathrm{Row}(\bar{Q}_k), \bar{C}_k - J] \text{ and } \bar{T}_k[\mathrm{Row}(\bar{T}_k), J] \text{ are nonsingular}\}, \tag{4.20}$$

since no cancellation occurs among terms with distinct J in the generalized Laplace expansion

$$\det \bar{D}_k = \sum_{J \in \mathcal{B}_k} \det \bar{Q}_k[\mathrm{Row}(\bar{Q}_k), \bar{C}_k - J] \cdot \det \bar{T}_k[\mathrm{Row}(\bar{T}_k), J].$$

Hence we obtain the following lemma concerning the nonzero fixed modes.

LEMMA 4.4. *Assume* (A1), (A2), *and* (A3) *as well as the nonsingularity of* $D(s)$ *of* (4.1). *The number of nonzero fixed modes is given by*

$$\sum_{k \in \bar{\Psi}_1} [\max\{\zeta(\bar{C}_k - J) + \xi_k(J) \mid J \in \mathcal{B}_k\} - \min\{\zeta(\bar{C}_k - J) + \eta_k(J) \mid J \in \mathcal{B}_k\}].$$

The combinatorial characterization of the nonexistence of fixed modes is now obtained, on the basis of which an efficient algorithm is designed in the next section.

THEOREM 4.2. *Assume* (A1), (A2), *and* (A3) *as well as the nonsingularity of* $D(s)$ *of* (4.1).

(1) $\lambda = 0$ *is not a fixed mode if and only if*

(C1) $D(0)$ *is nonsingular, i.e., there exist* $I \subseteq \mathrm{Row}(D)$ *and* $J \subseteq \mathrm{Col}(D)$ *such that* $Q^0[\mathrm{Row}(D) - I, \mathrm{Col}(D) - J]$ *and* $(T^0 + \hat{K})[I, J]$ *are nonsingular*;

(2) *The multiplicity of the zero fixed mode is given by*

$$\min\{\zeta(\bar{C} - J) + \eta(J) \mid J \in \mathcal{B}\} - \zeta(R)$$

where

$$\mathcal{B} = \{J \subseteq \bar{C} \mid \tilde{Q}[\mathrm{Row}(\tilde{Q}), \bar{C} - J] \text{ and } \tilde{T}[\mathrm{Row}(\tilde{T}), J] \text{ are nonsingular}\},$$

and $\eta(J)$ *is the lowest degree in* s *of a nonzero term in* $\det \tilde{T}[\mathrm{Row}(\tilde{T}), J]$ (*cf.* (4.8)).

(3) *There exist no nonzero fixed modes if and only if*

(C2) *For each* $k \in \bar{\Psi}_1$,

$$\max\{\zeta(\bar{C}_k - J) + \xi_k(J) \mid J \in \mathcal{B}_k\} = \min\{\zeta(\bar{C}_k - J) + \eta_k(J) \mid J \in \mathcal{B}_k\}.$$

Proof. (1) By Lemmas 4.2 and 2.1 applied to $D(0) = Q^0 + (T^0 + \hat{K})$ that is a mixed matrix with respect to $\mathbf{Q}$.

(2) See [26], or § 27 of [28].

(3) By Theorem 4.1, there exist no nonzero fixed modes if and only if $\deg_s \bar{\psi}_k(s) = 0$ for all $k \in \bar{\Psi}_1$ that is, in turn, equivalent to (C2) by (4.19). □

5. Algorithm of testing for the existence of fixed modes. In this section we describe an efficient algorithm for checking for the existence of fixed modes by means of Theorem 4.2. It is based on the algorithmic characterization of the irreducible components of the CCF of an LM-matrix (see § 22 of [28]) as well as on the fundamental facts concerning the independent-flow problem [13]. It would be interesting to compare this algorithm with those for testing the structural controllability (cf. [27] and § 29 of [28]) and for computing the dynamical degree (cf. [26] and § 27 of [28]).

First, note that (C1) in Theorem 4.2 can be checked readily by the efficient algorithm for computing the rank of a mixed matrix that uses arithmetic operations on rational numbers only [28], [31].

Before describing the concrete procedure for (C2), we will outline the basic idea in general terms. As shown in [28] and [32], the rank of an LM-matrix can be computed on the basis of Lemma 2.1 by finding a maximum independent flow [13] in a certain network and, moreover, the CCF can be obtained from the strong components [21] (or strongly connected components [1]) of the auxiliary network $\bar{N}$ associated with the maximum independent flow. By introducing appropriate costs associated with arcs, the quantities $\max\{\zeta(\bar{C}_k - J) + \xi_k(J)\}$ and $\min\{\zeta(\bar{C}_k - J) + \eta_k(J)\}$ appearing in (C2) are expressed as the maximum and the minimum of the cost of an independent flow in the strong component that corresponds to the block $\bar{D}_k$ of the CCF. If each arc in $\bar{N}$ is given the "length" $\bar{\gamma}$ that represents the imputed cost, then (C2) is equivalent to the graph-theoretic condition that there exists no directed cycle of nonzero length in the strong component. This condition is amenable to an efficient algorithm since, as noted in [27], this is equivalent to the existence of potentials associated with vertices such that the length of an arc is the difference of the potentials of the endvertices.

The concrete description of the algorithm for (C2) is as follows. It works with an auxiliary network $\bar{N} = (\bar{V}, \bar{A}; \bar{\gamma})$ with underlying graph $(\bar{V}, \bar{A})$ and length function $\bar{\gamma}$. Put $R = \mathrm{Row}\,(D)$, $C = \mathrm{Col}\,(D)$, and $V_Q = R_Q \cup C_Q$, $V_T = R_T \cup C_T$, where R_Q and R_T [respectively, C_Q and C_T] are disjoint copies of R [respectively, C]. Denote by $\varphi_Q : R \cup C \to R_Q \cup C_Q$ and $\varphi_T : R \cup C \to R_T \cup C_T$ the one-to-one correspondences. The vertex set $\bar{V}$ is defined as

$$\bar{V} = V_Q \cup V_T = (R_Q \cup C_Q) \cup (R_T \cup C_T). \tag{5.1}$$

The arc set $\bar{A}$ consists of five disjoint parts as follows:

$$\bar{A} = B_* \cup B^* \cup A_Q \cup A_T \cup A_M, \tag{5.2}$$

to be defined by (5.3), (5.4), (5.6), (5.7), and (5.8) below. The initial and terminal vertices of an arc $a \in \bar{A}$ are denoted, respectively, as $\partial^+ a$ and $\partial^- a$.

Denoting by $\hat{I} \subseteq R$ and $\hat{J} \subseteq C$ those subsets I and J that attain the maximum in Lemma 2.1 applied to $D(s)$ of (4.1), we put

$$B_* = \{(\varphi_T(i), \varphi_Q(i)) \mid i \in \hat{I}\} \cup \{(\varphi_T(j), \varphi_Q(j)) \mid j \in C - \hat{J}\}, \tag{5.3}$$

$$B^* = \{(\varphi_Q(i), \varphi_T(i)) \mid i \in R - \hat{I}\} \cup \{(\varphi_Q(j), \varphi_T(j)) \mid j \in \hat{J}\}. \tag{5.4}$$

Remark 5.1. Provided that $D(s)$ is nonsingular, both $Q_D(s)[R - \hat{I}, C - \hat{J}]$ and $T_D(s)[\hat{I}, \hat{J}]$ are nonsingular. Such a pair $(\hat{I}, \hat{J})$ can be found by an efficient algorithm using arithmetic operations in $\mathbf{Q}$ without involving s, since

$$\mathrm{rank}\, Q_D(s)[R - I, C - J] = \mathrm{rank}\, Q_D(1)[R - I, C - J], \qquad I \subseteq R, \quad J \subseteq C,$$

by (A2) and

$$\mathrm{rank}\, T_D(s)[I, J] = \text{term-rank}\, T_D(1)[I, J], \qquad I \subseteq R, \quad J \subseteq C,$$

by (A1) and (A3). See [28] and [31].

Let P be the pivotal transform of $Q \equiv Q_D(1) = Q^0 + Q^1$ with pivot $\hat{Q} \equiv Q[R-\hat{I}, C-\hat{J}]$, i.e., Row $(P) = (C-\hat{J}) \cup \hat{I}$, Col $(P) = (R-\hat{I}) \cup \hat{J}$, and

$$P = \begin{array}{c} \\ C-\hat{J} \\ \hat{I} \end{array} \begin{pmatrix} R-\hat{I} & \hat{J} \\ \hat{Q}^{-1} & \hat{Q}^{-1}Q[R-\hat{I}, \hat{J}] \\ -Q[\hat{I}, C-\hat{J}]\hat{Q}^{-1} & Q[\hat{I}, \hat{J}] - Q[\hat{I}, C-\hat{J}]\hat{Q}^{-1}Q[R-\hat{I}, \hat{J}] \end{pmatrix}. \tag{5.5}$$

With reference to P, we define

$$A_Q = \{(\varphi_Q(i), \varphi_Q(j)) \mid P_{ij} \neq 0,\ i \in (C-\hat{J}) \cup \hat{I},\ j \in (R-\hat{I}) \cup \hat{J}\}, \tag{5.6}$$

representing the linear dependence among the columns of the matrix $(I_d \mid Q_D)$ in (4.8).

The structure of T_D is represented by A_T and A_M. A_T has a one-to-one correspondence with the transcendental elements, i.e.,

$$A_T = \mathcal{S} \cup \mathcal{K} = \mathcal{N}(T^0) \cup \mathcal{N}(T^1) \cup \mathcal{N}(\hat{K}) \tag{5.7}$$

where we set $\partial^+ a = \varphi_T(i) \in R_T$ and $\partial^- a = \varphi_T(j) \in C_T$ if $a \in A_T$ is in the (i, j) entry of T_D. Note that parallel arcs can exist (e.g., if $T^0_{ij} T^1_{ij} \neq 0$). Since $T_D[\hat{I}, \hat{J}]$ is nonsingular, the bipartite graph $G_T = (R_T \cup C_T, A_T)$ with vertex set $R_T \cup C_T$ and arc set A_T has a matching $M (\subseteq A_T)$ such that $|M| = |\hat{I}| (= |\hat{J}|)$ and M covers $\hat{I}$ and $\hat{J}$ (i.e., $\hat{I} = \{\partial^+ a \mid a \in M\}$, $\hat{J} = \{\partial^- a \mid a \in M\}$). We define A_M as the set of reoriented arcs of M:

$$A_M = \{\bar{a} \mid a \in M\} \tag{5.8}$$

where $\bar{a}$ denotes the reorientation of a, i.e., $\partial^+ \bar{a} = \partial^- a$ and $\partial^- \bar{a} = \partial^+ a$.

The length $\bar{\gamma}: \bar{A} \to \mathbf{Z}$ is defined with reference to r_i and c_i $(i = 1, \cdots, d)$ of (4.4) as follows:

$$\begin{aligned}
&\bar{\gamma}(a) = r_i \quad \text{if } a = (\varphi_T(i), \varphi_Q(i)) \in B_*, \quad i \in \hat{I}, \\
&\bar{\gamma}(a) = c_j \quad \text{if } a = (\varphi_T(j), \varphi_Q(j)) \in B_*, \quad j \in C-\hat{J}, \\
&\bar{\gamma}(a) = -r_i \quad \text{if } a = (\varphi_Q(i), \varphi_T(i)) \in B^*, \quad i \in R-\hat{I}, \\
&\bar{\gamma}(a) = -c_j \quad \text{if } a = (\varphi_Q(j), \varphi_T(j)) \in B^*, \quad j \in \hat{J}, \\
&\bar{\gamma}(a) = 0 \quad \text{if } a \in A_Q, \\
&\bar{\gamma}(a) = 0 \quad \text{if } a \in \mathcal{N}(T^0) \cup \mathcal{N}(\hat{K}) \subseteq A_T, \\
&\bar{\gamma}(a) = 1 \quad \text{if } a \in \mathcal{N}(T^1) \subseteq A_T, \\
&\bar{\gamma}(a) = -\bar{\gamma}(a') \quad \text{if } a \in A_M \text{ is the reorientation of } a' \in M \subseteq A_T.
\end{aligned} \tag{5.9}$$

It may be noted here that $\bar{N}$ is defined with reference to a particular choice of $(\hat{I}, \hat{J})$ for the matrix $D(s)$, and without direct reference to $\bar{D}(s)$; however, the strong components of $\bar{N}$ are known to be determined independently of the choice of $(\hat{I}, \hat{J})$, and $\bar{D}(s)$ can be constructed easily from the strong components of $\bar{N}$.

Now we are ready to rephrase (C2) of Theorem 4.2 in terms of the network $\bar{N}$. For each strong component $\hat{G} = (\hat{V}, \hat{A})$ of $\bar{N}$ (where $\hat{V} \subseteq \bar{V}$, $\hat{A} \subseteq \bar{A}$), we consider the

condition that the sum of the lengths $\bar{\gamma}(a)$ along any directed cycle $\hat{C}$ in $\hat{G}$ is equal to zero, i.e.,

$$\sum_{a\in\hat{C}} \bar{\gamma}(a)=0 \quad (\forall \hat{C}\text{: directed cycle in } \hat{G}). \tag{5.10}$$

As observed in [27] and [28] (see also Remark 5.2 below), this condition is equivalent to the existence of a "potential" function $\pi: \hat{V}\to \mathbf{Z}$ such that

$$\bar{\gamma}(a)=\pi(\partial^- a)-\pi(\partial^+ a) \quad (\forall a\in \hat{A}). \tag{5.11}$$

THEOREM 5.1. *Assume* (A1), (A2), *and* (A3) *as well as the nonsingularity of* $D(s)$ *of* (4.1). *Condition* (C2) *in Theorem* 4.2 *holds if and only if*

(C3) *Each strong component of* $\bar{N}$ *either contains an arc of* $\mathcal{K}$ *or admits a potential function* π *such that* (5.11) *holds.*

Proof. First recall that the strong components on $\bar{N}$ correspond essentially to the irreducible diagonal blocks $\bar{D}_k$ of $\bar{D}$ of (4.10). As shown in § 4 (cf. (4.12)), we have

$$\det \bar{D}_k(s; 1, \cdots, 1)=\alpha_k s^{p_k}\cdot \bar{\psi}_k(s, \mathcal{S}, \mathcal{K})$$

with $\alpha_k\in\mathbf{Q}$, $p_k\in\mathbf{Z}$, and $\bar{\psi}_k(s, \mathcal{S}, \mathcal{K})\in \mathbf{Q}[s, \mathcal{S}, \mathcal{K}]-\mathbf{Q}[s]$ being irreducible. This polynomial $\bar{\psi}_k$ does not contain s, i.e., $\bar{\psi}_k\in\mathbf{Q}[\mathcal{S}, \mathcal{K}]$ if and only if the strong component corresponding to $\bar{D}_k$ satisfies (5.10). □

The overall computational complexity of testing (C1) and (C2), as well as testing for nonsingularity of $D(s)$, is dominated by that for the construction of the network $\bar{N}$ and is bounded by $O(d^3 \log d)$ in the worst case [9] and is usually much smaller than this worstcase bound (cf. [27] for the detail). Note also that the decomposition of $\bar{N}$ into strong components can be found [1] in time of $O(|\bar{A}|)$ and the potential function of (5.11) for a strong component $(\hat{V}, \hat{A})$, if any, can be found by constructing a spanning tree (cf. Remark 5.2) in time of $O(|\hat{A}|)$. It should be emphasized here that the whole algorithm involves only pivoting operations on the rational matrix Q_D whose entries are simple numbers such as ± 1 in practical applications.

Remark 5.2. The potential function π of (5.11) can be constructed as follows. First observe that, since $\hat{G}$ is strongly connected, (5.10) is equivalent to the condition that the length of a path depends only on the initial and terminal vertices of the path, i.e.,

$$\sum_{a\in P_1} \bar{\gamma}(a)=\sum_{a\in P_2} \bar{\gamma}(a)$$

if P_1 and P_2 are directed paths such that $\partial^+ P_1=\partial^+ P_2$ $(=u)$ and $\partial^- P_1=\partial^- P_2$ $(=v)$. In fact, taking a directed path P' with $\partial^+ P'=v$ and $\partial^- P'=u$ (such P' exists since $\hat{G}$ is strongly connected), we see

$$\sum_{a\in P_1} \bar{\gamma}(a)+\sum_{a\in P'} \bar{\gamma}(a)=0=\sum_{a\in P_2} \bar{\gamma}(a)+\sum_{a\in P'} \bar{\gamma}(a)$$

by (5.10). Hence the following procedure is valid: First fix a spanning tree $\hat{T}\subseteq \hat{A}$ and a vertex $u\in\hat{V}$; and for each $v\in\hat{V}$, set $\pi(v)$ equal to the length of the path in $\hat{T}$ connecting u to v; and finally check for the condition (5.11) for each $a\in\hat{A}-\hat{T}$.

Remark 5.3. The number of nonzero fixed modes can be computed efficiently based on Lemma 4.4 by solving the so-called independent-flow problem [13], [17] for each strong component of $\bar{N}$. The number of zero fixed modes can also be computed efficiently based on the formula of Theorem 4.2(2) using the matroid union/intersection algorithm.

Remark 5.4. In the particular case as treated in Pichai, Sezer, and Šiljak [33], the present characterization (i.e., Theorem 4.2 and Theorem 5.1) reduces to the graph-theoretic conditions given in Theorem 4 of [33], as follows.

When a system is described by (1.1) with "structured" matrices A, B, and C, having independent nonzero entries, and the feedback structure is specified by another "structured" matrix K, we have

$$Q_D = \begin{pmatrix} -sI & O & O \\ O & -I & O \\ O & O & -I \end{pmatrix}, \qquad T_D = \begin{pmatrix} A & B & O \\ O & O & K \\ C & O & O \end{pmatrix}$$

in (4.1) (see Remark 4.1). Note that the row-set R and the column-set C both correspond to the union of the sets X, U, and Y of variables $\mathbf{x}$, $\mathbf{u}$, and $\mathbf{y}$; let $\nu_R : R \to X \cup U \cup Y$ and $\nu_C : C \to X \cup U \cup Y$ denote the correspondence.

Then it is easy to see (by Linkage Lemma; cf., e.g., Proposition 7.1 of [28]) that (C1) in Theorem 4.2 is equivalent to (ii) in Theorem 4 of [33]. To define the network $\bar{N}$, we can choose $\hat{I} = \varnothing$, $\hat{J} = \varnothing$, since Q_D is nonsingular. The arc set of $\bar{N}$ consists of the following:

$$B_* = \{(\varphi_T(j), \varphi_Q(j)) \mid j \in C\},$$

$$B^* = \{(\varphi_Q(i), \varphi_T(i)) \mid i \in R\},$$

$$A_Q = \{(\varphi_Q(j), \varphi_Q(i)) \mid \nu_R(i) = \nu_C(j),\ i \in R, j \in C\},$$

$$A_T = \mathcal{N}(A) \cup \mathcal{N}(B) \cup \mathcal{N}(C) \cup \mathcal{N}(K),$$

$$A_M = \varnothing.$$

Note that B_*, B^*, and A_Q are in one-to-one correspondence to $X \cup U \cup Y$, and that cycles in $\bar{N}$ are in one-to-one correspondence to the cycles in the graph used in [33].

The matrix Q_D above satisfies (A2) with r_i and c_j in (4.4) defined as follows: $r_i = 1$ if $\nu_R(i) \in X$ and $=0$ otherwise; $c_j = 0$ for $j \in C$. Then the length $\bar{\gamma}(a)$ equals -1 if $a = (\varphi_Q(i), \varphi_T(i)) \in B^*$ and $\nu_R(i) \in X$, and vanishes otherwise. As easily seen, a strong component of $\bar{N}$ cannot admit a potential function π unless it consists of a single vertex. Therefore (i) in Theorem 4 of [33] implies (C3) in Theorem 5.1. Conversely, suppose (C3) holds together with (C1). Condition (C1) implies that the four vertices of $\bar{N}$ corresponding to one x-vertex are contained in a strong component of $\bar{N}$, which cannot admit a potential function, and therefore must contain an arc of $\mathcal{K}$. Thus, (C1) and (C3) together (in this special case) are equivalent to the two conditions given in Theorem 4 of [33]. See also Example 6.2.

6. Example.

Example 6.1. The algorithm described in § 5 as well as the derivation in § 4 is illustrated here by means of an example. Consider a matrix D of (4.1) (with $d = 9$)

given by

$$Q^0+sQ^1 = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & s & 0 & s & 0 \\ 1 & 0 & 1 & 0 & -1 & -s & s & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -s & 0 & -s & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

$$T^0+sT^1 = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & sf_1 & 0 & 0 & 0 & 0 & a_1 & 0 & 0 \\ 0 & a_2 & sf_2 & a_3 & a_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_7 & a_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_9 & sf_3 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

$$\hat{K} = \begin{array}{c} \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \\ w_7 \\ w_8 \\ w_9 \end{array} \begin{array}{c} \begin{array}{ccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & k_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & k_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Assumption (A2) is satisfied, where (4.4) holds true with

$$(r_1, \cdots, r_9) = (0, 0, 0, 0, 0, 0, 0, 1, 0),$$
$$(c_1, \cdots, c_9) = (0, 0, 0, 0, 0, -1, -1, -1, 1).$$

Note $\mathcal{S} = \{a_1, \cdots, a_9\} \cup \{f_1, \cdots, f_3\}$ and $\mathcal{K} = \{k_1, \cdots, k_4\}$.

By direct calculation we obtain

$$\det D = [s] \cdot [(a_9 - f_3)(f_1 f_2 s^2 - a_2)] \cdot [k_2(k_1 s + 1)(a_7 s - k_4 s + a_7 k_3 - a_6 k_4)]$$

where the brackets [ ] correspond to the three parts in (4.5). From this expression it follows by Lemma 4.1 that

$$\psi(s) = s \cdot (a_9 - f_3) \cdot (f_1 f_2 s^2 - a_2),$$

or

$$\psi(s) = s \cdot (f_1 f_2 s^2 - a_2),$$

since $(a_9 - f_3) \in \mathbf{C}$; note also that (4.7) holds with $\alpha = 1$ and $p = 1$.

The associated LM-matrix $\tilde{D}$ of (4.8) is given by

$$\tilde{D} = \left(\begin{array}{ccccccccc|ccccccccc}
w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \\
1 & & & & & & & & & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 1 & & & & & & & & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & 1 & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & 1 & & & & & & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & 1 & & & & & 0 & 0 & 0 & 0 & 1 & s & 0 & s & 0 \\
 & & & & & 1 & & & & 1 & 0 & 1 & 0 & -1 & -s & s & 0 & 0 \\
 & & & & & & 1 & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & 1 & & -s & 0 & -s & 0 & 0 & 0 & 0 & 0 & 1 \\
 & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
-t_1 & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & k_1 \\
 & -t_2 & & & & & & & & 0 & sf_1 & 0 & 0 & 0 & 0 & a_1 & 0 & 0 \\
 & & -t_3 & & & & & & & 0 & a_2 & sf_2 & a_3 & a_4 & 0 & 0 & 0 & 0 \\
 & & & -t_4 & & & & & & 0 & 0 & 0 & k_2 & 0 & 0 & 0 & 0 & 0 \\
 & & & & -t_5 & & & & & 0 & 0 & 0 & 0 & 0 & 0 & k_3 & 0 & a_5 \\
 & & & & & -t_6 & & & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_6 & 0 \\
 & & & & & & -t_7 & & & 0 & 0 & 0 & 0 & 0 & 0 & k_4 & a_7 & a_8 \\
 & & & & & & & -t_8 & & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & -t_9 & 0 & 0 & 0 & 0 & a_9 & sf_3 & 0 & 0 & 0
\end{array}\right).$$

By choosing, e.g.,

$$S = \begin{pmatrix}
1 & & & & & & & & \\
 & 1 & & & & & & & \\
 & & 1 & & & & & & \\
1 & & & 1 & & & & & \\
-1 & & & & 1 & 1 & & & \\
-1 & & & & & 1 & & & \\
 & & & & & & 1 & & \\
s & & & & & & & 1 & \\
 & & & & & & & & 1
\end{pmatrix}$$

in (2.3), we obtain the CCF $\bar{D}$ of (4.10):

$$\bar{D} = \left(\begin{array}{ccccccccccccccccc}
x_1 & w_2 & x_2 & x_3 & w_3 & x_4 & w_4 & x_5 & x_6 & w_9 & w_5 & w_6 & x_7 & x_8 & w_7 & w_1 & x_9 & w_8 \\
1 & & & 1 & & & & & & & & & & & & 1 & & \\
 & 1 & 0 & 1 & & & & & & & & & & & & & & \\
 & -t_2 & sf_1 & 0 & & & & & & & & & a_1 & & & & & \\
 & 0 & a_2 & sf_2 & -t_3 & a_3 & & a_4 & & & & & & & & & & \\
 & & & & 1 & & & & & & & & & & & & & \\
 & & & & & k_2 & -t_4 & & & & & & & & & & & \\
 & & & & & & 1 & & & & & & & & & 1 & & \\
 & & & & & & & -1 & -s & & & 1 & s & & & -1 & & \\
 & & & & & & & a_9 & sf_3 & -t_9 & & & & & & & & \\
 & & & & & & & & & 1 & & & & & & & & \\
 & & & & & & & & & & 1 & 1 & s & s & & -1 & & \\
 & & & & & & & & & & 0 & -t_6 & 0 & a_6 & & & & \\
 & & & & & & & & & & -t_5 & 0 & k_3 & 0 & & & a_5 & \\
 & & & & & & & & & & 0 & 0 & k_4 & a_7 & -t_7 & & a_8 & \\
 & & & & & & & & & & & & & & 1 & & & \\
 & & & & & & & & & & & & & & & s & 1 & 1 \\
 & & & & & & & & & & & & & & & -t_1 & k_1 & \\
 & & & & & & & & & & & & & & & & & -t_8
\end{array}\right).$$

The CCF $\bar{D}$ has 11 blocks with the following column sets: $\bar{C}_1 = \{x_1\}$, $\bar{C}_2 = \{w_2, x_2, x_3\}$, $\bar{C}_3 = \{w_3\}$, $\bar{C}_4 = \{x_4\}$, $\bar{C}_5 = \{w_4\}$, $\bar{C}_6 = \{x_5, x_6\}$, $\bar{C}_7 = \{w_9\}$, $\bar{C}_8 = \{w_5, w_6, x_7, x_8\}$, $\bar{C}_9 = \{w_7\}$, $\bar{C}_{10} = \{w_1, x_9\}$, $\bar{C}_{11} = \{w_8\}$.

The index sets of (4.13) are given by

$$\bar{\Psi}_1 = \{2, 6\}, \qquad \bar{\Psi}_2 = \{4, 8, 10\}$$

and $\bar{\psi}_k(s)$ of (4.12) for $k \in \bar{\Psi}_1 \cup \bar{\Psi}_2$ are as follows:

$$\bar{\psi}_2(s) = (f_1 f_2 s^2 - a_2), \qquad \bar{\psi}_6(s) = (a_9 - f_3),$$

$$\bar{\psi}_4(s) = k_2, \quad \bar{\psi}_8(s) = (a_7 s - k_4 s + a_7 k_3 - a_6 k_4), \quad \bar{\psi}_{10}(s) = (k_1 s + 1)$$

where note also that $\det \bar{D}_6 = s \cdot \bar{\psi}_6$.

We now apply the algorithm of § 5. Suppose we have found (cf. § 20 of [28])

$$\hat{I} = \{w_2, w_3, w_4, w_5, w_7, w_9\}, \qquad \hat{J} = \{x_2, x_3, x_4, x_6, x_7, x_8\},$$

and a matching

$$M = \{a_7, f_1, f_2, f_3, k_2, k_3\}.$$

Then we have the matrix P of (5.5) as follows:

$$P = \begin{array}{c} \\ x_1 \\ x_5 \\ x_9 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_7 \\ w_9 \end{array} \begin{array}{c} \begin{array}{cccccccccc} w_1 & w_6 & w_8 & x_2 & x_3 & x_4 & x_6 & x_7 & x_8 \end{array} \\ \left(\begin{array}{cccccccccc} 1 & & & & 1 & & & & \\ 1 & -1 & & & & & 1 & -1 & \\ 1 & & 1 & & & & & & \\ & & & & 1 & & & & \\ & & & & & & & & \\ 1 & & & & & & & & \\ -1 & 1 & & & & & & 1 & 1 \\ & & & & & & & & \\ & & & & & & & & \end{array} \right) \end{array}$$

and the auxiliary network $\bar{N}$ as depicted in Fig. 6.1; we write $x_i^T = \varphi_T(x_i)$, $x_i^Q = \varphi_Q(x_i)$, etc. In Fig. 6.1, the five subsets in (5.2) can be identified as follows: an arc belongs to B_* if it is of the form (w_i^T, w_i^Q) or (x_j^T, x_j^Q); to B^* if (w_i^Q, w_i^T) or (x_j^Q, x_j^T); to A_T if (w_i^T, x_j^T); to A_M if (x_j^T, w_i^T); and to A_Q otherwise. The associated length $\bar{\gamma}(a)$ of (5.9) is as follows: $\bar{\gamma}(a) = 1$ if

$$a \in \{(x_6^Q, x_6^T), (x_7^Q, x_7^T), (x_8^Q, x_8^T), (x_9^Q, x_9^T), (w_2^T, x_2^T), (w_3^T, x_3^T), (w_9^T, x_6^T)\};$$

$\bar{\gamma}(a) = -1$ if

$$a \in \{(w_8^Q, w_8^T), (x_2^T, w_2^T), (x_3^T, w_3^T), (x_6^T, w_9^T)\};$$

and $\bar{\gamma}(a) = 0$ otherwise.

The blocks in the CCF are determined from the strong components of $\bar{N}$. In particular, the strong component $\hat{G}_2$ consisting of $\{w_2^T, w_2^Q, w_3^T, x_2^T, x_3^T, x_3^Q\}$ and $\hat{G}_6$ of $\{w_9^T, x_5^T, x_5^Q, x_6^T, x_6^Q\}$ correspond to the diagonal blocks of $\{2, 6\} = \bar{\Psi}_1$. These two strong components are extracted in Fig. 6.2, where the length $\bar{\gamma}(a)$ is attached in parentheses to each arc a.

Theorem 5.1 reveals that $\hat{G}_2$ brings about nonzero fixed modes since it contains a directed cycle of nonzero length. On the other hand $\hat{G}_6$, having no directed cycle of nonzero length, introduces no nonzero fixed modes; $\hat{G}_6$ admits a potential function π: $\pi(x_6^T) = 1$, $\pi(w_9^T) = \pi(x_5^T) = \pi(x_5^Q) = \pi(x_6^Q) = 0$. We also see by Theorem 4.2(1) that $\lambda = 0$ is a fixed mode since $D(0)$ is singular; furthermore, by Theorem 4.2(2), $\lambda = 0$ is simple (i.e., with multiplicity one).

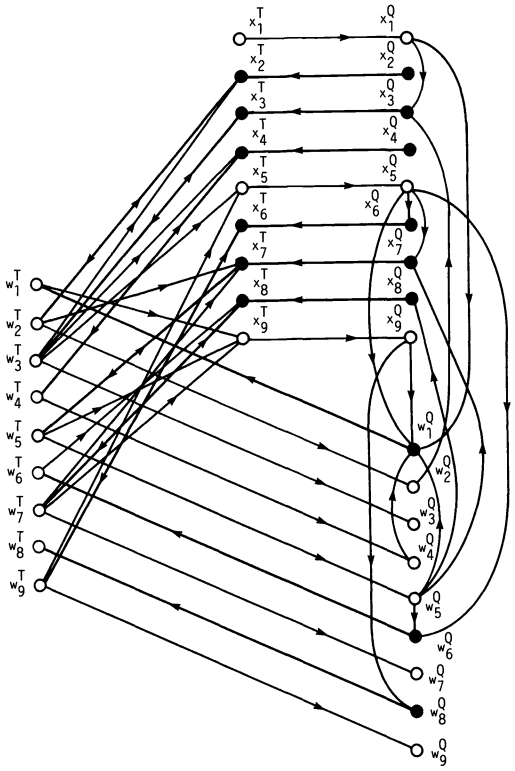

FIG. 6.1. *Auxiliary network* $\bar{N}$.

Example 6.2. Consider a scalar system in the state-space form (1.1) with $A=(0)$, $B=(b)$, $C=(c)$, and $K=(0)$. Obviously, this system has a simple fixed mode at $\lambda=0$, and no nonzero fixed modes. This fact is also revealed by Theorem 4.2(2) and (3). In particular, we see that $\lambda=0$ is a fixed mode since (C1) fails to hold, and that no nonzero fixed modes exist since (C2) (or equivalently (C3) in Theorem 5.1) holds true. In contrast, neither of the two conditions in Theorem 4 of Pichai, Sezer, and Šiljak [33] is satisfied. This shows that the conditions in [33] do not separate the existence of zero and nonzero fixed modes.

Example 6.3. Consider a decentralized system with three local stations described in the state space form (1.1) with

$$A=\begin{pmatrix} 0 & a_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ a_2 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B=\begin{pmatrix} 0 & 0 & 0 \\ b_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & b_3 \end{pmatrix},$$

and $C=I_6$, where a_1, a_2, b_1, b_2, b_3 are independent transcendentals; the feedback structure is specified by

$$K=\begin{pmatrix} k_1 & k_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & k_3 & k_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & k_5 & k_6 \end{pmatrix}.$$

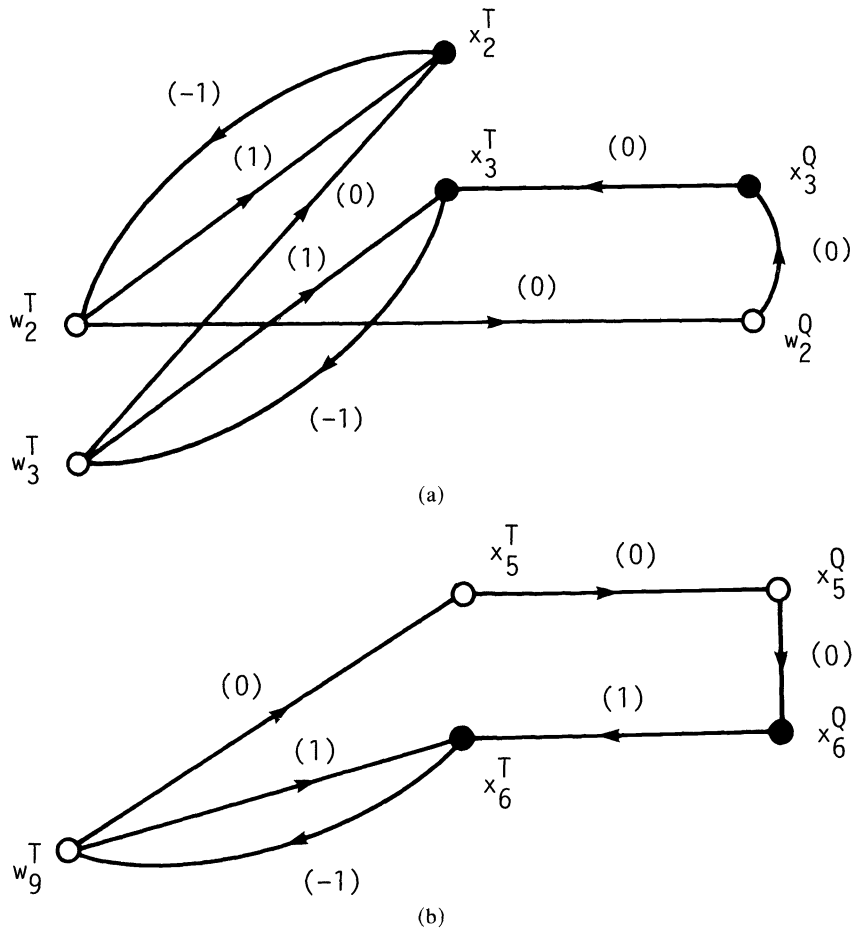

FIG. 6.2. *Strong components of* $\bar{\Psi}_1$. (a) *Strong component* $\hat{G}_2$. (b) *Strong component* $\hat{G}_6$.

(See Example 4 of Reinschke [33] and the references cited therein.)

The matrix Q_D in (4.1) is given by

$$Q_D = \begin{pmatrix} Q_A - sI_6 & O & O \\ O & -I_3 & O \\ I_6 & O & -I_6 \end{pmatrix}$$

where

$$Q_A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and satisfies (A2) with

$$(r_1, \cdots, r_{15}) = (1, 1, 1, 2, 1, 2; 0, 0, 0; 0, 0, 0, 1, 0, 1),$$
$$(c_1, \cdots, c_{15}) = (0, 0, 0, 1, 0, 1; 0, 0, 0; 0, 0, 0, 1, 0, 1)$$

Theorem 4.2 (as well as Theorem 5.1) reveals the fact that this system possesses the only fixed mode at $\lambda = 0$. The graph-theoretic criterion of Pichai, Sezer, and Šiljak [33] cannot detect this kind of fixed mode since $\lambda = 0$ would no longer be a fixed mode if the four unities in Q_A were replaced with independent variables.

7. Conclusion. The CCF of the matrix D of (3.2) provides a uniquely defined hierarchical decomposition of the set of variables $(\mathbf{x}, \mathbf{y}, \mathbf{u})$, where it is remembered that those variables correspond to the columns of the matrix D. The decomposition thus obtained is a generalization of the decomposition of the whole system into strongly connected subsystems that has played fundamental roles in various situations and is used, in particular, by Corfmat and Morse [8] in connection with the problem of pole assignment. In fact, when the system is in the usual state-space form (i.e., $F = I$, $G = I$, $H = I$ in (3.2)) and all the nonzero entries of A, B, C, and K are regarded as independent parameters (i.e., $Q_A = O$, $Q_B = O$, $Q_C = O$, and $Q_K = O$ in (3.4)), the decomposition of $(\mathbf{x}, \mathbf{y}, \mathbf{u})$ by means of the CCF agrees with the decomposition with respect to strong connectedness.

The hierarchical decomposition of $(\mathbf{x}, \mathbf{y}, \mathbf{u})$ induced by the CCF of D reveals which part of $\mathbf{x}$ depends on which part of the feedback loops. In particular, elimination of the feedback loops represented by those elements of $\mathcal{K}$ not contained in the diagonal blocks $\bar{D}_k$ in (4.10) never gives rise to fixed modes.

It is left for future investigation to clarify the practical significance of the decomposition by the CCF in relation to the pole assignment by dynamic compensation.

Acknowledgments. The author thanks Professor S. Shin of the University of Tsukuba for discussion. The comments of the anonymous referees were also helpful in revision.

REFERENCES

[1] A. V. Aho, J. E. Hopcroft, and J. D. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA, 1974.

[2] B. D. O. Anderson and D. J. Clements, *Algebraic characterization of fixed modes in decentralized control*, Automatica, 17 (1981), pp. 703-712.

[3] V. A. Armentano, *Eigenvalue placement for generalized linear systems*, Systems Control Lett., 4 (1984), pp. 199-202.

[4] ———, *The pencil $(sE - A)$ and controllability-observability for generalized linear systems: a geometric approach*, SIAM J. Control Optim., 24 (1986), pp. 616-638.

[5] K.-W. E. Chu, *A controllability condensed form and a state feedback assignment algorithm for descriptor systems*, IEEE Trans. Automat. Control, 33 (1988), pp. 366-370.

[6] D. Cobb, *Feedback and pole placement in descriptor variable systems*, Internat. J. Control, 33 (1981), pp. 1135-1146.

[7] ———, *Controllability, observability, and duality in singular systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 1076-1082.

[8] J. P. Corfmat and A. S. Morse, *Decentralized control of linear multivariable systems*, Automatica, 12 (1976), pp. 479-495.

[9] W. H. Cunningham, *Improved bounds for matroid partition and intersection algorithms*, SIAM J. Comput., 15 (1986), pp. 948-957.

[10] E. J. Davison, W. Gesing, and S. H. Wang, *An algorithm for obtaining the minimal realization of a linear time-invariant system and determining if a system is stabilizable-detectable*, IEEE Trans. Automat. Control, 23 (1978), pp. 1048-1054.

[11] A. L. Dulmage and N. S. Mendelsohn, *A structure theory of bipartite graphs of finite exterior dimension*, Trans. Roy. Soc. Canada, (3), 53 (1959), pp. 1-13.

[12] J. Edmonds, *Minimum partition of a matroid into independent subsets*, J. Nat. Bureau of Standards, 69B (1965), pp. 67-72.

[13] S. FUJISHIGE, *Algorithms for solving the independent-flow problems*, J. Oper. Res. Soc. Japan, 21 (1978), pp. 189-204.
[14] K. GLOVER AND L. M. SILVERMAN, *Characterization of structural controllability*, IEEE Trans. Automat. Control, 21 (1976), pp. 534-537.
[15] H. E. HUNTLEY, *Dimensional Analysis*, MacDonald, London, 1952.
[16] M. IRI, *Applications of matroid theory*, in Mathematical Programming—the State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 158-201.
[17] M. IRI AND S. FUJISHIGE, *Use of matroid theory in operations research, circuits and systems theory*, Internat. J. Systems Sci., 12 (1981), pp. 27-54.
[18] M. IRI AND N. TOMIZAWA, *A unifying approach to fundamental problems in network theory by means of matroids*, Electron. Comm. Japan, 58A (1975), pp. 28-35.
[19] H. KOBAYASHI AND T. YOSHIKAWA, *Graph-theoretic approach to controllability and localizability of decentralized control*, IEEE Trans. Automat. Control, 27 (1982), pp. 1096-1108.
[20] H. L. LANGHAAR, *Dimensional Analysis and Theory of Models*, John Wiley, New York, 1951.
[21] E. L. LAWLER, *Combinatorial Optimization: Networks and Matroids*, Holt, Reinhart and Winston, New York, 1976.
[22] C.-T. LIN, *Structural controllability*, IEEE Trans. Automat. Control, 19 (1974), pp. 201-208.
[23] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312-321.
[24] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473-480.
[25] H. MAEDA, *On structural controllability theorem*, IEEE Trans. Automat. Control, 26 (1981), pp. 795-798.
[26] K. MUROTA, *Use of the concept of physical dimensions in the structural approach to systems analysis*, Japan J. Appl. Math., 2 (1985), pp. 471-494.
[27] ———, *Refined study on structural controllability of descriptor systems by means of matroids*, SIAM J. Control Optim., 25 (1987), pp. 967-989.
[28] ———, *Systems Analysis by Graphs and Matroids—Structural Solvability and Controllability*, Springer-Verlag, Berlin, Heidelberg, 1987.
[29] ———, *On the irreducibility of layered mixed matrices*, Linear and Multilinear Algebra, 24 (1989).
[30] ———, *Some recent results in combinatorial approach to dynamical systems*, Linear Algebra Appl., 122/3 (1989).
[31] K. MUROTA AND M. IRI, *Structural solvability of systems of equations—a mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems*, Japan J. Appl. Math., 2 (1985), pp. 247-271.
[32] K. MUROTA, M. IRI, AND M. NAKAMURA, *Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of equations*, SIAM J. Algebraic Discrete Methods, 8 (1987), pp. 123-149.
[33] V. PICHAI, M. E. SEZER, AND D. D. ŠILJAK, *A graph-theoretic characterization of structurally fixed modes*, Automatica, 20 (1984), pp. 247-250.
[34] K. REINSCHKE, *Graph-theoretic characterization of fixed modes in centralized and decentralized control*, Internat. J. Control, 39 (1984), pp. 715-729.
[35] H. J. RYSER, *Indeterminates and incidence matrices*, Linear Multilinear Algebra, 1 (1973), pp. 149-157.
[36] M. E. SEZER AND D. D. ŠILJAK, *Structurally fixed modes*, Systems Control Lett., 1 (1981), pp. 60-64.
[37] R. W. SHIELDS AND J. B. PEARSON, *Structural controllability of multiinput linear systems*, IEEE Trans. Automat. Control, 21 (1976), pp. 203-212.
[38] G. C. VERGHESE, B. C. LÉVY, AND T. KAILATH, *A generalized state-space for singular systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 811-831.
[39] B. L. VAN DER WAERDEN, *Algebra*, Springer-Verlag, Berlin, 1955.
[40] S. H. WANG AND E. J. DAVISON, *On the stabilization of decentralized control systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 473-478.
[41] D. J. A. WELSH, *Matroid Theory*, Academic Press, London, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1403-1408, November 1989

010

A REMARK ON SIMULATED ANNEALING OF DIFFUSION PROCESSES*

G. ROYER†

Abstract. It is proved that simulated annealing for Kolmogorov processes takes place, as expected, for a cooling schedule corresponding to a fundamental constant of Wentzell and Freidlin.

Key words. diffusion, simulated annealing, large deviations

AMS(MOS) subject classifications. GOH10, GOJ70

1. Statement of the result. In their recent works Chiang, Hwang, and Sheu have proved convergence in law for diffusion processes Z_t on $\mathbb{R}^d$ ruled by

$$dZ_t = \varepsilon(t)\, dW_t - \nabla U(Z_t)\, dt, \tag{1}$$

$$\varepsilon(t) = \left(\frac{c}{\log(t)}\right)^{1/2} \qquad \text{(say for } t > 2\text{)} \tag{2}$$

toward a precise distribution supported by the set where the function U attains its global minimum. They get this result for $c > \frac{3}{2}\Lambda$, Λ being a fundamental nonnegative constant defined below, and conjectured that the condition $c > \Lambda$ is sufficient. In this note we prove this conjecture. Our method is just a patchwork combining estimates by Chiang, Hwang, and Sheu and by Gidas, Davies, and Simon.

We make the following assumptions:

(3_a) U is a C^2 function such that $U(\infty) = \infty$, $|\nabla U|(\infty) = \infty$, and $|\nabla U|^2 - \Delta U$ is bounded from below.

(3_b) The set of stationary points of U has a finite number of connected components.

Let us describe how Λ appears. When $\varepsilon > 0$ is fixed, the Kolmogorov process X_t^ε ruled by

$$dX_t^\varepsilon = \varepsilon\, dW_t - \nabla U(X_t^\varepsilon)\, dt \tag{4}$$

admit

$$\mu_\varepsilon = Z_\varepsilon^{-1} \exp(-2U/\varepsilon^2) \tag{5}$$

as invariant and reversible probability measure (the normalization constant Z is finite from hypothesis (3_a)). So we may view its transition semigroup P_t^ε as a semigroup of autoadjoint contractions in $L^2(\mu_\varepsilon)$. The corresponding infinitesimal generator is characterized by its action on test functions with compact support φ

$$L_\varepsilon(\varphi) = \frac{\varepsilon^2}{2}\Delta\varphi - \nabla U \cdot \nabla\varphi,$$

and the hypotheses on U ensures that its spectrum is discrete [7, Vol. IV, p. 120]. Let $\lambda(\varepsilon)$ be the first positive eigenvalue of L_ε and $\Lambda = \lim_{\varepsilon\to 0} -\varepsilon^2 \log(\lambda(\varepsilon))$.

Under the preceding hypothesis, Hwang and Sheu have established in [6], by purely probabilistic methods, the existence of $\Lambda = \lim_{\varepsilon\to 0} -\varepsilon^2 \log(\lambda(\varepsilon))$; they proved that $\Lambda \geqq 0$ and also give a formula for Λ. These results have their prototype in the book by Freidlin and Wentzell [3, p. 208].

* Received by the editors May 11, 1988; accepted for publication (in revised form) January 29, 1989.
† Département de Mathématiques et d'Informatique, Université d'Orléans, BP 6759, 45067 - Orleans Cedex 2, France.

Now we can state our slight improvement of the result of Chiang, Hwang, and Sheu. Let $\|\cdot\|$ be the total variation of a measure and let $P_{Z_t^x}$ be the law of a process satisfying (1), (2).

THEOREM 1.1. *If $c > \Lambda$, for any x, then $\|P_{Z_t^x} - \mu_{\varepsilon(t)}\|$ converges to zero when $t \to \infty$.*

We note that in most cases it is easy to prove the convergence of μ_ε when $\varepsilon \to 0$.

2. Proof. Let us sketch the strategy that we will follow. For annealing processes with finite or compact state space and discrete time, the annealing speed rate c (for a kind of convergence such as those in the preceding theorem) is directly linked with large deviations results for eigenvalues (see [4], [8]). On the other hand, the behavior of U at infinity is, to some extent, irrelevant for the calculus of Λ as well as c. In particular, in [1, p. 748], Theorem 1.1 is reduced to the "supernormal case," i.e., the case where $U(x) = |x|^4$ for $|x|$ greater than some fixed R_0 and (accessory) $\inf(U) = 0$. Fortunately, in this case some hypercontractive estimates of Nelson are available, under a suitable control when ε varies. We will therefore restrict ourselves to the supernormal case in the sequel. To begin, we adapt a lemma of Gidas [4] to a hypercontractive situation; we denote by Y_n a nonhomogenous Markov chain with arbitrary state space E and transition kernel N_n from n to $n+1$ (i.e., $E(f(Y_{n+1}) \mid Y_n) = (N_n f)(Y_n)$).

LEMMA 2.1. *Let (Y_n, N_n) defined for $n \geqq n_0$, obey the following hypotheses:*

(1°) *N_n admit reversible probabilities π_n and these π_n are mutually absolutely continuous.*

(2°) *For some constant K_n, for all measurable functions f on E,*

$$\|N_n f\|_{L^4(\pi_n)} \leqq K_n \|f\|_{L^2(\pi_n)}.$$

(3°) *Let $1 - r_n = \sup\{\|N_n f\|_2 / \|f\|_2;\ f \in L^2(\pi_n)$ and $\int f\, d\pi_n = 0\}$; then $\sum_{n \geqq n_0} r_n = +\infty$.*

(4°) *There exist constants $\gamma_n < 1$ and functions φ_n such that*

(a) $|\pi_{n+1} - \pi_n| \leqq \gamma_n \pi_n + \varphi_n \pi_{n+1}$, *as an inequality between measures,*

(b) $\gamma_n / r_n \to 0$,

(c) $K_n^2 \|\varphi_n\|_{L^2(\pi_n)} / r_n \to 0$.

(5°) *Y_n admit a density g_n with respect to π_n, and g_{n_0} is square-integrable.*

Then $\|P_{Y_n} - \pi_n\|$ converges to zero.

Proof. Suppose that $g_n \in L^2(\pi_n)$; then $N_n g_n$ is defined in $L^2(\pi_n)$ and from the reversibility of π_n, we easily get

$$g_{n+1}\pi_{n+1} = (N_n g_n)\pi_n. \tag{6}$$

Moreover, as $\gamma_n < 1$, $\pi_n \leqq ((1+\varphi_n)/(1-\gamma_n))\pi_{n+1}$. Let us denote in what follows by $\|\cdot\|_p$ the norm in $L^p(\pi_n)$. We have

$$\int g_{n+1}^2\, d\pi_{n+1} = \int (N_n g_n) g_{n+1}\, d\pi_n \leqq (1-\gamma_n)^{-1} \int (1+\varphi_n)(N_n g_n)^2\, d\pi_n$$

$$\leqq (1-\gamma_n)^{-1} \|1+\varphi_n\|_2 \|N_n g_n\|_4^2 \leqq (1-\gamma_n)^{-1} K_n^2 \|1+\varphi_n\|_2 \|g_n\|_2^2 < \infty.$$

Let $y_n = \int (1-g_n)^2\, d\pi_n = -1 + \int g_n^2\, d\pi_n$; from (6) we get

$$y_{n+1} = \int (1 - N_n g_n)^2\, d\pi_n + \int g_{n+1} N_n g_n (d\pi_n - d\pi_{n+1})$$

$$\leqq \int (N_n(1-g_n))^2\, d\pi_n + \gamma_n \int g_{n+1}^2\, d\pi_{n+1} + \int \varphi_n (N_n g_n)^2\, d\pi_n$$

$$\leqq (1-r_n)^2 y_n + \gamma_n (1 + y_{n+1}) + K_n^2 \|\varphi_n\|_2 (1 + y_n)$$

and finally

$$y_{n+1} \leqq a_n y_n + b_n, \tag{7}$$

where

$$a_n(1-\gamma_n) = (1-r_n)^2 + K_n^2 \|\varphi_n\|_2, \quad \text{and} \quad b_n(1-\gamma_n) = \gamma_n + K_n^2 \|\varphi_n\|_2.$$

Hypothesis (3°) yields $\prod_{n \geqq n_0} a_n = 0$ and $b_n/(1-a_n) \to 0$, and from (7) we easily deduce $y_n \to 0$ and a fortiori the lemma. □

To use this lemma, we will consider a process obtained by replacing in (1) the function $\varepsilon(t)$ by a stepwise function that is constant on each interval of some sequence $[t_n, t_{n+1}[$ and that takes there the value $e(n) = \varepsilon(t_n)$ (sometimes denoted also e_n below). It happens that the simple choice $t_n = n$ leads to processes that diverge too much from Z_t; instead we use $t_n = \sum_{k=1}^{n-1} k^{-\alpha}$, where $0 < \alpha < 1$. Later the exponent α will be chosen small enough. We have

$$e_n^{-2} \sim \left(\frac{1-\alpha}{c}\right) \log(n). \tag{8}$$

Precisely, we define inductively a process Y_n, as follows, for n no less than an n_0 to be fixed later; for Y_{n+1} we take the value at a certain time τ_n of the Kolmogorov process $X_t^{e(n)}$ with initial value Y_n (and naturally we impose that the Wiener process of (4) be independent of the Y_k, $k \leqq n$). The value of τ_n is chosen to easily compare Z_{t_n} to Y_n: as in [1], by the time change $t \to \beta(n, t)$ determined by

$$\int_{t_n}^{\beta(n,t)} \varepsilon^2(u)/e^2(n)\, du = t,$$

any annealing process (1) is transformed after t_n into $Z_{\beta(n,t)}$ owning the same martingale component (in law) $e(n) W_t$ as $X_t^{e(n)}$. So to compare the transition from Z_{t_n} to $Z_{t_{n+1}}$ to the transition from Y_n to Y_{n+1} we set $\beta(n, \tau_n) = t_{n+1}$, i.e., $\tau_n = \int_{t_n}^{t_{n+1}} \log(t_n)/\log(u)\, du$; it is easily seen that $\tau_n \sim n^{-\alpha}$. Now, we study why Lemma 2.1 applies to our Y_n. Hypothesis (1°) is satisfied with $\pi_n = \mu_{e(n)}$ given by (5), and $N_n = P_{\tau_n}^{e(n)}$. Using $P_t^\varepsilon = \exp(tL_\varepsilon)$, we see that $r_n = 1 - \exp(-\tau_n \exp((-\Lambda + \rho(n))/e_n^2))$, where $\lim \rho(n) = 0$; thus,

$$r_n \sim n^{-1+(1-\alpha)(1-\Lambda/c)} \tag{9}$$

and condition (3°) reduces to $c > \Lambda$.

Explicit constants for hypercontractivity are established in [2]. We now deduce K_n from this article; let ψ_ε be the groundstate of the Schrödinger operator $-\Delta + V_\varepsilon$ in $L^2(dx)$ that is unitarily equivalent to $-\underline{L}_\varepsilon = -2\varepsilon^{-2} L_\varepsilon = -\Delta + 2\varepsilon^{-2} \nabla U \cdot \nabla$. We know (see [2] that ψ_ε^2 is the density of μ_ε and that $V_\varepsilon = \varepsilon^{-4} |\nabla U|^2 - \varepsilon^{-2} \Delta U$.

LEMMA 2.2. *For some constant k_0 and small enough positive ε and δ, in the operator sense* $-2 \log(\psi_\varepsilon) \leqq \delta(-\Delta + V_\varepsilon) + k_0 \varepsilon^{-2} + (\varepsilon^2/4\delta^2)$.

Proof. We have $-2 \log(\psi_\varepsilon) = 2U\varepsilon^{-2} + \log(Z_\varepsilon)$; as $\varepsilon^2 \log(Z_\varepsilon) \to 0$ when $\varepsilon \to 0$ (recall that $\inf(U) = 0$), the corresponding term can be absorbed by $k\varepsilon^{-2}$.

$$2U \leqq \delta(\varepsilon^{-2} |\nabla U|^2 - \Delta U) + k_1 + \varepsilon^4/4\delta^2. \tag{10}$$

For $|x| \geqq R_0$, we have

$$U(x) = |x|^4, \quad \nabla U(x) = 4x|x|^2, \quad \Delta U(x) = 12|x|^2,$$

and for small δ we may use $2U(x) + \delta \Delta U(x) \leqq 4|x|^4$.

If $|x| \geqq \sup(R_0, \varepsilon(4\delta)^{-1/2})$, $4|x|^4 \leqq \delta\varepsilon^{-2} |\nabla U|^2(x)$ and (10) is valid.

If $R_0 \leqq |x| \leqq \varepsilon(4\delta)^{-1/2}$, $4|x|^4 \leqq \varepsilon^4/4\delta^2$ and (10) is also verified.

Finally, we can choose k_1 large enough to majorize $2U - \delta \Delta U$ for $|x|$ smaller than R_0. □

Lemma 2.2 provides us with the necessary estimate to apply Theorem 4.2 via Theorems 4.6 and 5.1 of [2]. We get that if

$$t = \int_2^4 \left(\frac{\delta}{p}\right) dp = \delta \log(2), \tag{11}$$

then

$$\|e^{-tL_\varepsilon} f\|_4 \leqq e^M \|f\|_2 \tag{12}$$

with

$$M = \int_2^4 2bp^{-2}\, dp = \frac{b}{2} \quad \text{and} \quad b = A - \frac{d \log(\delta)}{4} + k_0 \varepsilon^{-2} + \frac{\varepsilon^2}{4\delta^2}. \tag{13}$$

This last formula corresponds to (5.2) in [2], where the constant c can be taken to be zero, because $-L_\varepsilon$ is positive in $L^2(\mu_\varepsilon)$. (A is a constant depending only on the dimension d.) To get a hypercontractive estimate for N_n, we must use (12) for $t = \tau_n e_n^2/2$, so we choose $\delta = \tau_n e_n^2/2 \log(2)$, so (12), (13) gives a constant K_n such that

$$\log(K_n) \sim k \log(n) n^{2\alpha} \quad \text{with } k = (1-\alpha)\frac{\log^2(2)}{2c}. \tag{14}$$

To verify hypothesis (4°), we recall the formula

$$\frac{\partial}{\partial \beta} \nu_\beta(x) = -(U(x) - \bar{U}_\beta)\nu_\beta(x) \tag{15}$$

for the density of probability $\nu_\beta(x) = Z_\beta^{-1} \exp(-\beta U(x))$, where $\bar{U}_\beta$ is the mean of U with respect to ν_β, and β is a positive number (identified with $2\varepsilon^{-2}$ below). We denote by H_n the compact set $\{U \leqq n^{3\alpha}\}$. As $\bar{U}_\beta$ vanishes when β goes to infinity, using (15) we see that, when restricted to the exterior of H_n, π_{n+1} is smaller than π_n for n large; so (4°)(a) holds on $H_n^c = \mathbb{R}^d - H_n$ if we let φ_n be the restriction of the density π_n/π_{n+1} to H_n^c; then elementary Laplace theory yields $\log \|\varphi_n\|_{L^2(\pi_n)} \sim -((1-\alpha)/c) \log(n) n^{3\alpha}$. This estimate, joined to (9) and (14) gives (4°)(c). Let us study (4°)(a) on H_n; let x belong to H_n and $\beta_n = 2e_n^{-2}$; from (15) we get, if n is large enough to ensure $\bar{U}_\beta \leqq n^{3\alpha}$ for any $\beta \geqq \beta_n$:

$$|\nu_{\beta_{n+1}}(x) - \nu_{\beta_n}(x)| \leqq n^{3\alpha}(\beta_{n+1} - \beta_n) \sup\{\nu_\beta(x);\ \beta_n < \beta < \beta_{n+1}\}.$$

On the other hand, for any x and $\beta_n \leqq \beta \leqq \beta_{n+1}$ $\nu_\beta(x)/\nu_{\beta_n}(x) \leqq \exp((\beta_{n+1} - \beta_n) n^{3\alpha})/\nu_{\beta_n}(H_n)$; since $\beta_{n+1} - \beta_n \sim 2(1-\alpha)/cn$ this last quantity converges to one if $\alpha < \frac{1}{3}$. We select such an α and we see that (4°)(a) holds with $\gamma_n = n^{3\alpha}(\beta_{n+1} - \beta_n)$. And the hypothesis (4°)(b) is verified if we choose a small α, due to the estimate (9).

Finally, n_0 will be chosen according to all the conditions "n large enough" that appear above and in the sequel. For Y_{n_0} we take any variable whose law is the normalized restriction of the law of $Z(t_{n_0})$ to a ball; all hypothesis of Lemma 2.1 are fulfilled and therefore $\|P^n_{Y_{n_0}} - \pi_n\|$ converges to zero. To prove the theorem, we must estimate $\Delta_n = \|P_{Z(t_n)} - P_{Y_n}\|$. To begin with, $\varepsilon > 0$ being given, we may realize $\Delta_{n_0} \leqq \varepsilon$ when we build Y_{n_0}. For a given n, let us consider a process $\hat{Z}_t$, $t \geqq t_n$, with the same equation as Z_t but the same law at time t_n as Y_n. Clearly the law at time t_{n+1} of $\hat{Z}$ and Z will differ of a quantity less than Δ_n; thus

$$\Delta_{n+1} \leqq \Delta_n + d_{n+1} \quad \text{where } d_{n+1} = \|P_{Y_{n+1}} - P_{\hat{Z}(t_{n+1})}\|. \tag{16}$$

To bound d_{n+1}, we must compare at time τ_n two processes with the same initial law P_{Y_n}: $X_t^{e(n)}$ that obey (4), and $\bar{Z}_t := \hat{Z}_{\beta(n,t)}$; the time change β has been defined so that

$$d\bar{Z}_t = e(n)\, d\bar{W}_t - \nabla U(\bar{Z}_t) e^2(n) \varepsilon^{-2}(\beta(n,t))\, dt. \tag{17}$$

Let us first study how the law of X_t concentrates. From the proof of Lemma 2.1, we know that the density g_n has a norm bounded in L^2 by a constant a: so we have $P(Y_n \in K_p^c) \leqslant a(\pi_n(K_p^c))^{1/2}$ for any compact $K_p = \{U \leqslant p\}$, and the last quantity is for $c' > c(1-\alpha)^{-1}$ and n large majorized by $n^{-2p/c'}$; as the exponent $1+(\alpha/2)$ reappears below we call it s and we fix p so that

$$P(Y_n \in K_p^c) \leqq n^{-s}. \tag{18}$$

Let T_r be the exit time of the open ball B_r centered at zero. For a Kolmogorov process with parameter $e(n)$, starting from a point $x \in K_p$, we have for r suitably chosen and n large

$$P_x^{e(n)}(T_r \leqq 1) \leqq n^{-s} \quad \text{for } x \in K_p \tag{19}$$

it suffices to apply a result of Freidlin and Wentzell [3, p. 105], since the action

$$S = \inf\left\{\frac{1}{2}\int_0^1 |\dot{\varphi}(s) + \nabla U(\varphi(s))|^2\, ds\right\},$$

$$\varphi,\ \varphi(0) \in K_p, \quad \exists s \in [0,1] |\varphi(s)| \geqq r$$

can be made as large as we want (it is no less than $\inf_{x \in K_p, y \in B_r^c}\{2(U(y) - U(x))\}$).

Let χ be the indicator function of the event $\{|X_s| < r \text{ for } 0 \leqq s \leqq \tau_n\}$, ($X_s^{e(n)}$ is shortened to X_s), and let $\bar{\chi}$ be the analogous function relative to $\bar{Z}$. We have the Girsanov formula

$$E(\bar{\chi} f(\bar{Z}_{\tau_n}) - \chi f(X_{\tau_n})) = E((\varphi - 1)\chi f(X_{\tau_n})). \tag{20}$$

Because χ is present, we can modify U into $\bar{U}$ outside B_r to get a function with gradient bounded by M, and use as φ the martingale $\varphi = \exp(A - B/2)$ with

$$A = -\int_0^{\tau_n} e^{-1}(n) \nabla \bar{U}(X_u) q(u)\, dW_u, \qquad B = \int_0^{\tau_n} e^{-2}(n) |\nabla \bar{U}|^2(X_u) q^2(u)\, du,$$

$$q(u) = e^2(n)\varepsilon^{-2}((\beta(n,u))) - 1 = (\log(\beta(n,u))/\log(t_n)) - 1.$$

By a martingale argument we get (see [1] for details) $E(|\varphi - 1|) \leqq E^{1/2}((\varphi-1)^2) \leqq \|\exp(B) - 1\|_\infty^{1/2}$, and we have

$$c\|B\|_\infty \leqq M^2 \log(n) \int_0^{\tau_n} q^2(u)\, du \quad \text{(taking } \beta(n,u) = \beta \text{ as variable)}$$

$$= M^2 \int_{t_n}^{t_{n+1}} (\log(\beta) - \log(t_n))^2/\log(\beta)\, d\beta$$

$$\leqq M^2 (t_{n+1} - t_n)^3/3t_n^2 \log(t_n) \sim M^2(1-\alpha) n^{-2-\alpha}/3\log(n)$$

so

$$E(|\varphi - 1|) \leqq n^{-s} \quad \text{for large } n. \tag{21}$$

Now we apply (20), when $f = 1$, to the process conditioned to start at $x \in K_p$; due to estimate (19) we have $E(\chi) \geqq 1 - n^{-s}$ and thus $E(\bar{\chi}) \geqq 1 - 2n^{-s}$; and (20), (21) give $\|P_{\bar{Z}_{\tau(n)^x}} - P_{X_{\tau(n)^x}}\| \leqq 4n^{-s}$. Actually, since the starting distribution verifies only (18) we

have $d_n \leqq 6n^{-s}$, and (16) gives $\Delta_n \leqq \varepsilon + \sum_{n_0}^{n} 6k^{-s}$. Also, if n_0 is suitably chosen, $\Delta_n \leqq 2\varepsilon$ for $n \geqq n_0$; this proves the theorem since ε is arbitrary (we have only proved the convergence of the subsequence t_n, but a last step from t_n to t, $t \in [t_n, t_{n+1}[$ can be treated similarly).

Acknowledgment. I thank the referee for his effective help in rectifying the first version of this article.

REFERENCES

[1] T. S. CHIANG, C. R. HWANG, AND S. J. SHEU, *Diffusion for global optimization in* $\mathbb{R}^n$, SIAM J. Control Optim., 25 (1987), pp. 737-753.

[2] E. B. DAVIES AND B. SIMON, *Ultracontractivity and the heat kernel for Schrödinger operators*, J. Funct. Anal., 59 (1984), pp. 335-395.

[3] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[4] B. GIDAS, *Global minimization via the Langevin equation*, in Proc. 24th IEEE Conference on Decision and Control.

[5] C. R. HWANG AND S. J. SHEU, *Large time behaviors of perturbed diffusion Markov processes*, preprint.

[6] ———, *The asymptotic behavior of the second eigenvalue of perturbed Fokker-Planck operators*, preprint.

[7] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1978.

[8] R. HOLLEY AND D. STROOCK, in Conference at "Colloque P. Levy," June 1987.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1409-1439, November 1989

011

ERGODIC THEOREMS FOR DISCRETE TIME STOCHASTIC SYSTEMS USING A STOCHASTIC LYAPUNOV FUNCTION*

SEAN P. MEYN†

Abstract. Sufficient conditions are established under which the law of large numbers and related ergodic theorems hold for nonlinear stochastic systems operating under feedback. It is shown that these conditions hold whenever a moment condition is satisfied; this may be interpreted as a generalizaton of the martingale property.

If, in addition, a stochastic controllability condition holds, then it is shown that the underlying distributions governing the system converge to an invariant probability at a geometric rate.

The key assumption used is that a Markov chain with stationary transition probabilities exists that serves as a state process for the closed loop system.

Key words. nonlinear systems, stochastic systems, Lyapunov functions, Markov chains

AMS(MOS) subject classifications. 60J10, 60J20, 93E15

1. Introduction.

In this paper we study the asymptotic behavior of discrete time nonlinear stochastic systems under feedback. Our principal assumption is that a Markov chain with stationary transition probabilities $\mathbf{\Phi}$ exists that may serve as a state process for the closed loop system.

In a large number of applications $\mathbf{\Phi}$ evolves on a subset $\mathbf{X} \subset \mathbb{R}^n$, and is generated by a nonlinear difference equation

$$\Phi_{k+1} = F(\Phi_k, w_{k+1}), \qquad k \in \mathbb{Z}_+ \tag{1}$$

where the disturbance $\mathbf{w}$ is an independent and indentically distributed (i.i.d.) process on $\mathbb{R}^p$. Under the appropriate smoothness conditions on the function F and the distribution of $\mathbf{w}$, it has been shown in Meyn and Caines (1988) that the asymptotic behavior of $\mathbf{\Phi}$ is determined by invariant probabilities on $\mathbf{X}$ whenever a crude stability condition is satisfied, and the *weak stochastic controllability* condition holds. In this case there exist probabilities $\{\pi_x, \tilde{\pi}_x : x \in \mathbf{X}\}$ such that for every initial condition $\Phi_0 = x \in \mathbf{X}$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) = \int f \, d\tilde{\pi}_x \quad \text{a.s. } [P_x], \tag{2}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} E_x[g(\Phi_k)] = \int g \, d\pi_x \tag{3}$$

for a large class of functions f and g on $\mathbf{X}$. When (1) is viewed as a deterministic input/output (i.o.) system with input $\mathbf{w}$ and output $\mathbf{\Phi}$, weak stochastic controllability is equivalent to the forward accessibility criterion used in nonlinear system theory (see Jakubczyk and Sontag (1988)).

It is worth noting that the left-hand side of (2) will be in general a random variable, and hence $\tilde{\pi}_x$ is a *random probability* (i.e., a countably additive function from $\mathcal{B}(\mathbf{X})$ to $L^1(\mathbf{X}^{Z_+}, \mathcal{B}(\mathbf{X}^{Z_+}), P_x)$ with the properties $\tilde{\pi}\{A\} \geqq 0$ almost surely (a.s.) $[P_x]$ for $A \in \mathcal{B}(\mathbf{X})$, and $\tilde{\pi}\{\mathbf{X}\} = 1$ a.s. $[P_x]$). With one or two exceptions (that will be brought to the attention of the reader where they occur), in the present paper only ordinary (deterministic) probabilities will be considered.

* Received by the editors June 13, 1988; accepted for publication (in revised form) January 30, 1989.

† This work was conducted at the Department of Systems Engineering, the Australian National University. Present address, University of Illinois, Coordinated Science Laboratory, 1101 W. Springfield Avenue, Urbana, Illinois 61801.

In the continuous time case, Φ is typically a diffusion process and the theory of stochastic Lyapunov functions has been a successful tool for assessing its stability properties (see Kushner (1967), Kushner (1972), and Has'minskiĭ (1980)). The idea is that if a positive function $V: \mathbf{X} \to \mathbb{R}_+$ exists such that $V(\Phi_k)$ is a super martingale and hence decreases in an average sense, then under general conditions Φ will converge to a level set of the function V with probability one.

It is reasonable to expect that stochastic Lyapunov functions should be a useful tool in the discrete time case as well, and this is indeed the case (see, for example, Kushner (1967), Solo (1978), and Goodwin, Ramadge, and Caines (1981)). However when Φ is weakly stochastically controllable, the existence of a stochastic Lyapunov function is all but ruled out. In many cases (for example, when V is a quadratic) the level sets of V are sets of Lebesgue measure zero, yet weak stochastic controllability implies that the set of limit points of the sequence

$$\{\Phi_k : k \in \mathbb{Z}_+\} \tag{4}$$

has nonempty interior with probability one for all initial conditions. In fact, for almost every sample path, the set of limit points of (4) is equal to the support of the probability $\tilde{\pi}_x$, that always has nonempty interior under the weak stochastic controllability hypothesis.

In this paper we examine an alternative stochastic Lyapunov function that is perfectly compatible with weak stochastic controllability, and that always exists for stable, linear systems even when conventional stochastic Lyapunov functions do not exist. If Φ is weakly stochastically controllable, then the existence of this Lyapunov function may be used to prove that (3) holds at an exponential rate for a large class of functions g, and will allow us to establish generalizations of (2) and (3) even when the weak stochastic controllability condition is not satisfied.

The existence of the limit (2) for general functions on $\mathbf{X}$ is closely connected to a condition called Harris recurrence (see § 2). The following result is taken from Meyn and Caines (1988) (for the "if" part) and Athreya and Ney (1980) (for the converse) and is valid for general Markov chains on a general state space.

PROPOSITION 1.1. *The Markov chain Φ is positive Harris recurrent if and only if a unique invariant probability π exists, and* (2) *holds with $\tilde{\pi}_x = \pi$ for every positive Borel function $f: \mathbf{X} \to \mathbb{R}_+$ and every initial condition $x \in \mathbf{X}$.*

Hence if Φ is not positive Harris recurrent then there will exist functions on $\mathbf{X}$ for which (2) fails to hold, and it is natural to search for a restricted class of functions for which the law of large numbers does hold.

Take, for example, a Markov chain defined by the linear model

$$\Phi_{k+1} = A\Phi_k + Bw_{k+1} \tag{5}$$

where A and B are, respectively, $n \times n$ and $n \times p$ matrices, $\mathbf{w} = \{w_k : k \geqq 1\}$ is an i.i.d. Gaussian stochastic process on $\mathbb{R}^p$ with $w_k \sim N(0, I)$ for all k, and the deterministic initial condition $\Phi_0 \in \mathbb{R}^n$ is given.

Suppose that the eigenvalues of A fall strictly within the unit disk in $\mathbb{C}$. In this case a unique invariant probability π exists that is supported on the controllability subspace $L \subset \mathbb{R}^n$. Hence if the pair (A, B) is not controllable, then π is supported on a subspace of $\mathbb{R}^n$ whose dimension is strictly less than n. In the case where the matrix A is full rank and the initial condition lies in the complement of the set L, the process Φ will approach the set L at a geometric rate, but never reach it. From these observations it is obvious that the law of large numbers cannot hold for general measurable functions on $\mathbb{R}^n$ when the initial condition lies outside of the set L. Take, for example, $f = \mathbf{1}_L$,

the indicator function of L. In this case the average on the left side of the equality in (2) is equal to zero for each $N \in \mathbb{Z}_+$, and the right-hand side of this equality is equal to one. However it may be verified that (2) will hold for continuous functions (whenever the right-hand side is meaningful) regardless of the controllability of (A, B).

In the theory of recurrent Markov chains, the set L^c would be regarded simply as a set of measure zero on which the process $\mathbf{\Phi}$ is transient. In cases where $\mathbf{\Phi}$ does not satisfy a recurrence condition, some other assumption must be used to connect points in the state space together, and in this paper we accomplish this by supposing that $\mathbf{\Phi}$ evolves on a metric space and that the Feller property holds. When a certain Lyapunov function exists, a complete generalization of the linear example will be established.

2. Preliminaries. Let $\mathbf{X}$ be a locally compact separable metric space. We let $\mathbf{C}$ denote the set of bounded and continuous functions $f: \mathbf{X} \to \mathbb{R}$, and $\mathcal{M}$ the set of probabilities on $\mathcal{B}(\mathbf{X})$, the Borel field on $\mathbf{X}$. A sequence $\{\mu_k : k \in \mathbb{Z}_+\} \subset \mathcal{M}$ of probabilities converges *weakly* to $\mu_\infty \in \mathcal{M}$ if

$$\lim_{k\to\infty} \int f\, d\mu_k = \int f\, d\mu_\infty \tag{6}$$

for all $f \in \mathbf{C}$, and this will be denoted $\mu_k \xrightarrow{\text{weakly}} \mu_\infty$ as $k \to \infty$. It is well known (see Billingsley (1968)) that $\mathcal{M}$ is a metrizable topological space and that a subset $\mathcal{A} \subset \mathcal{M}$ is precompact if and only if it is tight, i.e., for all $\varepsilon > 0$ there exists a compact set $K \subset \mathbf{X}$ such that

$$\mu\{K\} \geqq 1 - \varepsilon, \qquad \mu \in \mathcal{A}.$$

A function $V: \mathbf{X} \to \mathbb{R}_+$ is called a *moment* if there exists a sequence of compact sets $K_n \uparrow \mathbf{X}$ such that

$$\lim_{n\to\infty} \inf_{x \in K_n^c} V(x) = \infty$$

where we adopt the convention that the infimum of a function over the empty set is infinity. It is easily verified that $\mathcal{A} \subset \mathcal{M}$ is tight if and only if a moment V exists such that

$$\sup_{\mu \in \mathcal{A}} \int V\, d\mu < \infty.$$

In Theorem 6.6 of Parthasarathy (1967) it is shown that there exists a sequence of uniformly continuous functions $\{g_n : n \in \mathbb{Z}_+\} \subset \mathbf{C}$ with the property that

$$\mu_k \xrightarrow{\text{weakly}} \mu_\infty \quad \Leftrightarrow \quad \forall n \in \mathbb{Z}_+, \qquad \lim_{k\to\infty} \int g_n\, d\mu_k = \int g_n\, d\mu_\infty. \tag{7}$$

We let P denote a Feller Markov transition function on $(\mathbf{X}, \mathcal{B}(\mathbf{X}))$. That is, for all $x \in \mathbf{X}$, $A \in \mathcal{B}(\mathbf{X})$, and $f \in \mathbf{C}$,

$$P(x, \cdot) \text{ is a probability on } \mathcal{B}(\mathbf{X}),$$

$$P(\cdot, A) \text{ is } \mathcal{B}(\mathbf{X})\text{-measurable},$$

$$\int f(y) P(\cdot, dy) \text{ is continuous.}$$

The k-fold iterates of P are defined inductively by $P^1 \triangleq P$, and

$$P^{k+1}(x, A) \triangleq \int P(x, dy) P^k(y, A),$$

and for $f \in \mathbf{C}$ and $\mu \in \mathcal{M}$ we use the standard notation

$$P^k f(\cdot) \triangleq \int P^k(\cdot, dy) f(y), \qquad \mu P^k(\cdot) \triangleq \int \mu(dx) P^k(x, \cdot).$$

The majority of results in the theory of Markov chains on general state spaces, as well as the results to be presented here, require the following hypothesis (see, for example, Orey (1971) and Nummelin (1984)). We say Φ is *irreducible* if the following condition is satisfied. The event $\{\Phi \text{ enters } A\} \triangleq \cup_{k=0}^{\infty} \{\Phi_k \in A\}$, and the event $\{\Phi \in A \text{ i.o.}\} \triangleq \cap_{N=0}^{\infty} \cup_{k=N}^{\infty} \{\Phi_k \in A\}$.

Irreducibility Hypothesis. There exists a set $A \in \mathcal{B}(\mathbf{X})$, an integer n_0, a number $\lambda_0 > 0$, and a probability φ, such that

(i) $P_x\{\Phi \text{ enters } A\} > 0$ for all $x \in \mathbf{X}$;

(ii) $\sum_{k=1}^{n_0} P_x\{\Phi_k \in E\} \geqq \lambda_0 \varphi\{E\}$ for all $x \in A$, and $E \in \mathcal{B}(\mathbf{X})$.

Φ is called *Harris recurrent* if the following condition from Athreya and Ney (1980) is satisfied.

Recurrence Hypothesis. Φ satisfies the irreducibility hypothesis, and for every $x \in \mathbf{X}$,

$$P_x\{\Phi \text{ enters } A\} = 1.$$

A subset $B \subset \mathbf{X}$ is called *absorbing* if $P(x, B) = 1$ for every $x \in B$. If B is absorbing, then the Markov chain Φ may be restricted to the set B, and B is called a *Harris set* if the restricted process is Harris recurrent.

When Φ is irreducible, the set A used in the irreducibility hypothesis will be called *petite.* It may be verified that if Φ is Harris recurrent and the set A is petite, then it satisfies the smallness condition introduced in Nummelin and Tuominen (1982).

Many of the important limit theorems for Markov chains require the existence of an invariant measure. That is, a σ-finite measure π on $\mathcal{B}(\mathbf{X})$ with the property

$$\pi\{A\} = \int \pi(dx) P(x, A) \quad \text{for all } A \in \mathcal{B}(\mathbf{X}).$$

It is shown in Nummelin (1984), and Orey (1971) that if the recurrence hypothesis holds, then an essentially unique invariant measure π exists. If the invariant measure is finite, then it may be normalized to a probability measure and in this case Φ is called *positive Harris recurrent.*

If Φ is irreducible, then there exists an integer $m \in \mathbb{Z}_+$ called the *period* of Φ, and a collection of sets $\{E_1, \cdots, E_m\}$ with the property that

$$P\mathbf{1}_{E_{i+1}} = \mathbf{1}_{E_i} \quad \text{and} \quad P^m \mathbf{1}_{E_i} = \mathbf{1}_{E_i}$$

for each $1 \leqq i \leqq m$. If Φ is Harris recurrent with invariant measure π, then $\pi\{(\cup E_i)^c\} = 0$.

The following proposition shows that if Φ is positive Harris recurrent and aperiodic ($m = 1$), then its underlying distributions converge to the invariant probability for all initial conditions. If in addition, the distribution of the hitting time $\tau_A \triangleq \min\{k \geqq 1 : \Phi_k \in A\}$ to a petite set A possesses geometrically decaying tails, uniformly for initial conditions lying in A, then the underlying distributions of Φ converge to π at a geometric rate.

Define the total variation norm $\|\mu - \nu\|$ for $\nu, \mu \in \mathcal{M}$ by

$$\|\mu - \nu\| \triangleq \sup \left| \int f d\mu - \int f d\nu \right|$$

where the supremum is taken over all Borel functions $f : \mathbf{X} \to [-1, 1]$.

PROPOSITION 2.1. *Suppose that Φ is positive Harris recurrent and aperiodic with invariant probability π. Then,*

(i) *For each initial condition distribution $\mu \in \mathcal{M}$,*

$$\lim_{k\to\infty} \|\mu P^k - \pi\| = 0.$$

(ii) *If in addition the set $A \in \mathcal{B}(\mathbf{X})$ defined in the recurrence hypothesis satisfies*

$$\sup_{x\in A} E_x[r^{\tau_A}] < \infty, \tag{8}$$

for some $r > 1$, then there exists $\rho < 1$, and an extended real-valued function $M \in L^1(\mathbf{X}, \mathcal{B}(\mathbf{X}), \pi)$ such that for each $x \in \mathbf{X}$,

$$\|\delta_x P^k - \pi\| = \|P^k(x, \cdot) - \pi(\cdot)\| \leqq M(x)\rho^k, \qquad k \in \mathbb{Z}_+.$$

(iii) *If the conditions of* (ii) *hold, and an initial condition distribution $\mu_0 \in \mathcal{M}$ satisfies*

$$E_{\mu_0}[r^{\tau_A}] < \infty,$$

then there exists $\rho < 1$ and $M < \infty$ such that

$$\|\mu_0 P^k - \pi\| \leqq M\rho^k, \qquad k \in \mathbb{Z}_+.$$

(iv) *Suppose that the conditions of* (iii) *hold, and $f: \mathbf{X} \to \mathbb{R}$ satisfies*

$$\sup_{k\in\mathbb{Z}_+} E_{\mu_0}[|f(\Phi_k)|^{1+\delta}] < \infty$$

for some $\delta > 0$. Then there exists $\rho < 1$ and $M < \infty$ such that

$$\left| E_{\mu_0}[f(\Phi_k)] - \int f\, d\pi \right| \leqq M\rho^k, \qquad k \in \mathbb{Z}_+.$$

For a proof of Proposition 2.1(i) see Nummelin (1984). Results (ii) and (iii) may be found in Nummelin and Tuominen (1982) and result (iv) follows from (iii) and Hölder's inequality.

An aperiodic positive Harris recurrent Markov chain is sometimes called *ergodic.* If for all $x \in \mathbf{X}$ there exists $\rho(x) < 1$ and $M(x) < \infty$ such that

$$\|P^k(x, \cdot) - \pi(\cdot)\| \leqq M\rho^k, \qquad k \in \mathbb{Z}_+,$$

then Φ will be called *geometrically ergodic.* This is weaker then the notion of geometric ergodicity introduced in Tweedie (1983), and stronger than that of Nummelin (1984). In § 3 we present sufficient conditions for geometric ergodicity using a stochastic Lyapunov function.

The following stability conditions will be shown to be closely connected to Harris (respectively, positive Harris) recurrence:

(S1) For each initial condition $x \in \mathbf{X}$ and each $\varepsilon > 0$, there exists a compact subset $K \subset \mathbf{X}$ such that

$$P_x\{\Phi \in K \text{ i.o.}\} = \lim_{k\to\infty} P_x\left\{ \bigcup_{i=k}^{\infty} \{\Phi_i \in K\} \right\} \geqq 1 - \varepsilon.$$

(S2) For each initial condition $x \in \mathbf{X}$ and each $\varepsilon > 0$ there exists a compact subset $K \subset \mathbf{X}$ such that

$$\liminf_{k\to\infty} P_x\{\Phi_k \in K\} \geqq 1 - \varepsilon.$$

It may be shown that if a moment V exists such that

$$\liminf_{k\to\infty} V(\Phi_k) < \infty \quad \text{a.s. } [P_x]$$

for each $x \in \mathbf{X}$ then condition (S1) holds, and if

$$\limsup_{k\to\infty} E_x[V(\Phi_k)] < \infty$$

for each $x \in \mathbf{X}$ then condition (S2) is satisfied.

It is evident that condition (S2) implies condition (S1). In Rosenblatt (1971) a strengthening of condition (S1) is used to establish the existence of a σ-finite invariant measure for Feller Markov chains. Condition (S2) is called *boundedness in probability* in Meyn and Caines (1988), and is simply the tightness hypothesis of Billingsley (1968). In Beneš (1968) a similar condition (among other assumptions) is used to establish the existence of an invariant probability for a continuous time Feller Markov process, and under the assumptions already made on $\mathbf{\Phi}$, condition (S2) implies the existence of an invariant probability (see Foguel (1969)).

The following is a simple result, but it appears to be new.

PROPOSITION 2.2. *Suppose that $\mathbf{\Phi}$ satisfies the irreducibility condition with an open petite set A. Then $\mathbf{\Phi}$ is Harris recurrent if and only if condition* (S1) *is satisfied, and $\mathbf{\Phi}$ is positive Harris recurrent if and only if condition* (S2) *holds.*

Similar results may be found in Tuominen and Tweedie (1979). See, in particular, Proposition 3.5 or Theorem 7.1 of that paper. Theorem 7.1 gives sufficient conditions for a generalization of Harris and positive Harris recurrence. In this result it is assumed that the state space is compact, which is much stronger than condition (S2). On the other hand, the conditions that $\mathbf{\Phi}$ is Feller and an open petite set exist are stronger than the remaining hypotheses of Theorem 7.1(i) and (ii).

The proof of Proposition 2.2 will be given below. It would be very useful if Proposition 2.2 held without the assumption that the set A is open. However, this is not the case as can be seen from the following simple example. Let $\mathbf{X} = \mathbb{R}$, and consider the Markov transition function P defined by $P(0, \{0\}) = 1$, and

$$P(x, \{1/2x\}) = 2^{-|x|} \quad \text{for } x \in \mathbb{R},$$
$$P(x, \{0\}) = 1 - 2^{-|x|}, \qquad x \neq 0.$$

The corresponding Markov chain $\mathbf{\Phi}$ has the Feller property, satisfies condition (S2), and satisfies the irreducibility condition with $A = \{0\}$ and $\varphi = \delta_0$. However, $\mathbf{\Phi}$ is *not* Harris recurrent since for any $x \in \mathbf{X}$, $x \neq 0$,

$$P_x\{\Phi_k \neq 0 \text{ for all } k\} = 2^{-|x|}2^{-|x/2|}2^{-|x/4|} \cdots$$
$$= 2^{-2|x|} > 0.$$

Let $\Lambda : \mathbf{X} \times \mathcal{B}(\mathbf{X}) \to [0, 1]$ denote the function

$$\Lambda(x, A) \triangleq P_x\{\mathbf{\Phi} \text{ enters } A\} = P_x\left\{ \bigcup_{k=0}^{\infty} \{\Phi_k \in A\} \right\}. \tag{9}$$

The following result is taken from Orey (1971) and Lemma 4.1 of Cogburn (1975).

LEMMA 2.1. *Let $\mathbf{\Phi}$ be a Feller Markov chain on $\mathbf{X}$, and $B \in \mathcal{B}(\mathbf{X})$.*

(i) *For each $x \in \mathbf{X}$, $\lim_{k\to\infty} \Lambda(\Phi_k, B) = \mathbf{1}_{\{\mathbf{\Phi} \in B \text{ i.o.}\}}$ a.s. $[P_x]$.*

(ii) *If B is open then for each $k \in \mathbb{Z}_+$ the functions*

$$\Lambda(\cdot, B) \quad \text{and} \quad P^k(\cdot, B)$$

are lower semicontinuous.

One important consequence of Lemma 2.1(i) is that if the recurrence hypothesis holds, then $P_x\{\mathbf{\Phi} \in B \text{ i.o.}\} = 1$ for every set $B \in \mathcal{B}(\mathbf{X})$ of positive φ-measure, and every initial condition $x \in \mathbf{X}$. This is called φ-recurrence in Orey (1971) and is, in fact, the usual definition of Harris recurrence (see Nummelin (1984) for further details).

LEMMA 2.2. *Suppose that the irreducibility hypothesis is satisfied with an open petite set A. Then we have the following*:

(i) *If* $K \in \mathcal{B}(\mathbf{X})$ *has positive* φ*-measure, then*

$$\Lambda(x, K) > 0 \quad \text{for all } x \in \mathbf{X}.$$

(ii) *If K is compact, then there exists* $T_0 \in \mathbb{Z}_+$ *such that for all* $x \in K$,

$$\sum_{k=1}^{T_0} P^k(x, A) \geqq \frac{1}{T_0} \quad \text{and} \quad \sum_{k=1}^{2T_0} P^k(x, \cdot) \geqq \frac{\lambda_0}{T_0^2} \varphi\{\cdot\}.$$

Lemma 2.2 shows that when an open petite set exists, all compact sets of positive φ-measure are also petite. Conversely, in the common case where $P: \mathbf{C}_0 \to \mathbf{C}_0$, where $\mathbf{C}_0 \subset \mathbf{C}$ denotes the set of continuous functions on $\mathbf{X}$ that vanish at infinity, it may be shown that every petite set is precompact.

Proof of Lemma 2.2. Result (i) follows immediately from the irreducibility hypothesis.

By Lemma 2.1 and the irreducibility hypothesis, the sets

$$O_n \triangleq \left\{ x \in \mathbf{X}: \sum_{k=1}^{n} P^k(x, A) > \frac{1}{n} \right\}$$

form an open cover of K, and by compactness there exists $T_0 \in \mathbb{Z}_+$ such that $\sum_{k=1}^{T_0} P^k(x, A) > 1/T_0$ for all $x \in K$. This establishes the first inequality in (ii).

The second inequality follows from the first by integrating over the set A and using the irreducibility hypothesis. Assume that T_0 is so large that

$$\sum_{j=1}^{T_0} P^j(y, \cdot) \geqq \lambda_0 \mathbf{1}_{y \in A} \phi(\cdot), \qquad y \in \mathbf{X}.$$

Then for every $B \in \mathcal{B}(\mathbf{X})$ and $x \in K$,

$$\begin{aligned} \sum_{k=1}^{2T_0} P^k(x, B) &\geqq \frac{1}{T_0} \left(\sum_{i=0}^{T_0} P^i \right) \left(\sum_{j=1}^{T_0} P^j \right)(x, B) \\ &\geqq \frac{\lambda_0}{T_0} \left(\sum_{i=0}^{T_0} P^i \right)(x, A) \varphi\{B\} \\ &\geqq \left(\frac{\lambda_0}{T_0^2} \right) \varphi\{B\}. \end{aligned}$$

□

Proof of Proposition 2.2. If $\mathbf{\Phi}$ is Harris recurrent, then for all compact sets K that have positive φ-measure

$$P_x\{\mathbf{\Phi} \in K \text{ i.o.}\} = 1$$

for all $x \in \mathbf{X}$. This follows from the remark below Lemma 2.1. Hence Harris recurrence implies condition (S1).

If $\mathbf{\Phi}$ is positive Harris recurrent, then for each initial condition $x \in \mathbf{X}$ the invariant probability π may be decomposed

$$\pi = \frac{1}{\lambda} \sum_{k=1}^{\lambda} \nu_k^x$$

where $\lambda \in \mathbb{Z}_+$, $\{\nu_k^x: 1 \leqq k \leqq \lambda\}$ are probabilities on $\mathcal{B}(\mathbf{X})$, and for every set $B \in \mathcal{B}(\mathbf{X})$,

$$\lim_{k \to \infty} P_x\{\Phi_{k\lambda + i} \in B\} = \nu_i^x\{B\}.$$

It follows that for all compact sets K,

$$\liminf_{k \to \infty} P_x\{\Phi_k \in K\} \geqq \min_{1 \leqq i \leqq \lambda} \nu_i^x\{K\}$$

and this shows that positive Harris recurrence implies (S2).

We now show that (S1) implies Harris recurrence under the conditions of Proposition 2.2. Let $x \in \mathbf{X}$, $\varepsilon > 0$, and choose a compact set $K \subset \mathbf{X}$ such that

$$P_x\{\Phi \in K \text{ i.o.}\} \geqq 1 - \varepsilon. \tag{10}$$

By the irreducibility condition $\Lambda(x, A) > 0$ for all $x \in \mathbf{X}$, and by lower semicontinuity, the sets

$$\mathcal{O}_\delta \triangleq \{x \in \mathbf{X}: \Lambda(x, A) > \delta\}$$

form an open cover of K. By compactness there exists $\delta > 0$ such that

$$\Lambda(x, A) \geqq \delta \quad \text{for all } x \in K. \tag{11}$$

From this, Lemma 2.1, and (10) it follows that

$$\begin{aligned} P_x\{\lim_{k\to\infty} \Lambda(\Phi_k, A) = 1\} &= P_x\{\limsup_{k\to\infty} \Lambda(\Phi_k, A) \geqq \delta\} \\ &\geqq P_x\{\Phi \in K \text{ i.o.}\} \geqq 1 - \varepsilon. \end{aligned}$$

Since ε is arbitrary it follows that

$$\mathbf{1}_{\{\Phi \in A \text{ i.o.}\}} = \lim_{k\to\infty} \Lambda(\Phi_k, A) = 1 \quad \text{a.s. } [P_x],$$

and hence Φ is Harris recurrent.

If condition (S2) holds then since this implies condition (S1), Φ is Harris recurrent. Since an invariant probability exists when (S2) holds it follows that Φ is positive Harris recurrent. □

We now describe how the irreducibility condition may be established using ideas from nonlinear control theory. For precise definitions and results see § 5, and Meyn and Caines (1988).

For a Markov chain of the form (1), suppose that $\mathbf{X}$ is a smooth n-dimensional manifold, and that the disturbance process $\mathbf{w}$ evolves on an open set $\mathcal{O}_w \subset \mathbb{R}^p$. We call a point $y \in \mathbf{X}$ *attainable* from $x \in \mathbf{X}$ if for some $T \in \mathbb{Z}_+$, and some sequence $(w_1^*, \cdots, w_T^*) \in \mathcal{O}_w^T$,

$$\Phi_T = y \quad \text{when } \Phi_0 = x \text{ and } (w_1, \cdots, w_T) = (w_1^*, \cdots, w_T^*).$$

If for each $x \in \mathbf{X}$, the set of all attainable points from x has nonempty interior, then Φ is called *forward accessible* (see Jakubczyk and Sontag (1990)).

Suppose that the function $F: \mathbf{X} \times \mathbb{R}^p \to \mathbf{X}$ defined in (1) is C^∞, and that the distribution of the disturbance process $\mathbf{w}$ possesses a lower semicontinuous density p_w with $\mathcal{O}_w = \{x \in \mathbb{R}^p : p_w(x) > 0\}$. Under these conditions, the Markov chain Φ is called *weakly stochastically controllable* if it is forward accessible.

It is shown in Meyn and Caines (1988) that if Φ is weakly stochastically controllable, for every $x \in \mathbf{X}$ there exists an open set A_x containing x, and a probability φ_x such that condition (ii) of the irreducibility hypothesis holds.

Suppose now that there exists a distinguished point $x_0 \in \mathbf{X}$ such that for every initial condition $x \in \mathbf{X}$ and every $\varepsilon > 0$, there exists $T_0 \in \mathbb{Z}_+$ and some sequence $(w_1^*, \cdots, w_\infty^*) \in \mathcal{O}_w^\infty$ such that

$$|\Phi_T - x_0| < \varepsilon \quad \text{when } \Phi_0 = x \text{ and } (w_1, \cdots, w_T) = (w_1^*, \cdots, w_T^*), \qquad T \geqq T_0.$$

In this case Φ is called *asymptotically controllable* to x_0. This is actually much weaker than the standard definition but is sufficient for our purposes.

If Φ is weakly stochastically controllable and there exists a point $x_0 \in \mathbf{X}$ such that Φ is asymptotically controllable to x_0, then the irreducibility condition is satisfied with $A = A_{x_0}$ and $\varphi = \varphi_{x_0}$, and Φ is aperiodic.

3. Stochastic Lyapunov functions. In this section we show how stochastic Lyapunov functions may be used to verify the geometric ergodicity conditions of Proposition 2.1, and to hence establish the exponential asymptotic stability of the flow of distributions governing the process $\mathbf{\Phi}$.

Let $V:\mathbf{X}\to\mathbb{R}_+$ be a positive measurable function. The following drift condition was introduced in Kushner (1967) to compute estimates of the probability that a Markov chain will leave a compact set before a given finite time (see Theorem 3, p. 86).

For some $0<\lambda<1$, $K>0$, and all $x\in\mathbf{X}$,

$$PV(x)\leqq\lambda V(x)+K. \tag{12}$$

In terms of the stochastic process $\mathbf{\Phi}$ this condition may be expressed as

$$E[V(\Phi_{k+1})\,|\,\mathcal{F}_k]\leqq\lambda V(\Phi_k)+K$$

where $\mathcal{F}_k$ is the sigma algebra generated by past and present values of $\mathbf{\Phi}$:

$$\mathcal{F}_k\triangleq\sigma\{\Phi_0,\cdots,\Phi_k\}. \tag{13}$$

There is a great deal of motivation for introducing this condition. First, it is a natural generalization of the martingale property: in the degenerate case where $\lambda=1$ and $K=0$ (12) becomes

$$E[V(\Phi_{k+1})\,|\,\mathcal{F}_k]\leqq V(\Phi_k). \tag{14}$$

Hence in this case $(V(\Phi_k),\mathcal{F}_k)$ is a positive supermartingale, and by the martingale convergence theorem (see Doob (1953)) for each initial condition $x\in\mathbf{X}$ there exists a random variable $V_\infty=V_\infty(x)$ such that

$$\lim_{k\to\infty}V(\Phi_k)=V_\infty\quad\text{a.s.}\,[P_x],$$

and furthermore $E_x[V_\infty]\leqq V(x)<\infty$.

A positive function $V:\mathbf{X}\to\mathbb{R}_+$ satisfying (14) is called a *super harmonic* function in Nummelin (1984) and Revuz (1975), and a *stochastic Lyapunov function* in the stochastic systems theory literature.

As mentioned in the Introduction, if a Lyapunov function V satisfying (14) exists whose level sets have zero Lebesgue measure, then $\mathbf{\Phi}$ cannot be weakly stochastically controllable. Furthermore, if the irreducibility condition holds, then (14) implies that the irreducibility measure φ must be singular with respect to the Lebesgue measure, and this is ruled out in a large number of examples. On the other hand, (12) is extremely useful when the irreducibility condition is satisfied.

PROPOSITION 3.1. *Suppose that V is a continuous moment satisfying* (12), *and $\mathbf{\Phi}$ is a Feller Markov chain. If the irreducibility hypothesis holds for an open set $A\subset\mathbf{X}$, and if $\mathbf{\Phi}$ is aperiodic, then $\mathbf{\Phi}$ is geometrically ergodic.*

Similar results may be found in Nummelin (1984) and Nummelin and Tuominen (1982), and the proof will closely follow Theorem 3.1 of that paper. The principal difference between our result and these lies in our consideration of the topology of the state space. Our result also differs in the form of the "test function" V.

A connected result may also be found in the dissertation of Chan (1986) for a Markov chain of the form (1). However, this result relies on an exponential stability hypothesis on the deterministic dynamical system obtained when the disturbance $\mathbf{w}$ is set equal to zero, and a global Lipschitz condition on the function F.

The original inspiration for such test function methods for establishing positive recurrence lies in the paper by Foster (1953).

The hypothesis that V is a continuous moment is not crucial. This property is assumed because it implies that the set $K_t \triangleq \{x \in \mathbf{X}: V(x) \leqq t\}$ is compact, and hence by Lemma 2.2 petite for all t sufficiently large. If this property can be established by some other means for an arbitrary positive measurable function V, then the conclusions of Proposition 3.1 still hold. In fact, in this case the Feller property is no longer needed.

We have recently discovered that, under the conditions of Proposition 3.1, the Central Limit Theorem holds for measurable functions whose square is dominated by V. This result will appear in a sequel to this paper.

The following result exhibits the close relationship between the existence of a function satisfying (12), and the tail of the distribution of the first entrance time to a compact set.

LEMMA 3.1. (i) *Suppose that* $V: \mathbf{X} \to \mathbb{R}_+$ *is a positive measurable function satisfying* (12), *and let* $r \in \mathbb{R}_+$ *satisfy* $1 > r^{-1} > \lambda$. *Then for each* $t > K/(r^{-1} - \lambda)$, *there exists a constant* $B_1 > 0$ *such that for every* $x \in \mathbf{X}$,

$$E_x[r^{\tau_{K_t}}] \leqq B_1 V(x) + B_1.$$

(ii) *Let* $r \geqq 1$ *and* $A \in \mathcal{B}(\mathbf{X})$. *Then with*

$$V(\cdot) \triangleq \mathbf{1}_{A^c}(\cdot) E_{(\cdot)}[r^{\tau_A}]$$

we have

$$PV \leqq r^{-1} V \quad \text{on } A^c$$

and hence (12) *holds if* PV *is uniformly bounded on* A.

Proof. The proof of (ii) follows along the same lines as the proof of Proposition 6.1 of Tweedie (1975).

To prove (i) observe that by (12) we have for every $x \in \mathbf{X}$,

$$rPV(x) \leqq V(x) - ((1 - \lambda r) V(x) - rK). \tag{15}$$

Let $U(x) \triangleq (1 - \lambda r) V(x) - rK$ for $x \in \mathbf{X}$, $\varepsilon \triangleq (1 - \lambda r) t - rK > 0$, and observe that by the conditions of the lemma,

$$\mathbf{1}_{K_t^c}(x) U(x) \geqq \varepsilon \mathbf{1}_{K_t^c}(x) \quad \text{for all } x \in \mathbf{X}. \tag{16}$$

Equation (15) implies that for all $x \in \mathbf{X}$,

$$rP\mathbf{1}_{K_t^c} V(x) \leqq V(x) - U(x) \tag{17}$$

where the operator "$rP\mathbf{1}_{K_t^c}$" is defined for a positive function $f: \mathbf{X} \to \mathbb{R}_+$ by

$$rP\mathbf{1}_{K_t^c} f(x) \triangleq r \int_{K_t^c} P(x, dy) f(y).$$

Applying this operator to both sides of (17) gives

$$\begin{aligned} (rP\mathbf{1}_{K_t^c})^2 V &\leqq V - U - rP\mathbf{1}_{K_t^c} U \\ &\leqq V - U - \varepsilon rP\mathbf{1}_{K_t^c} 1, \end{aligned}$$

and by induction we have for all $x \in \mathbf{X}$,

$$\varepsilon \sum_{k=1}^{\infty} (rP\mathbf{1}_{K_t^c})^k 1(x) \leqq V(x) - U(x). \tag{18}$$

The left-hand side of this equation may be transformed as follows:

$$\begin{aligned}\sum_{k=1}^{\infty}(rP\mathbf{1}_{K_t^c})^k 1(x) &= \sum_{k=1}^{\infty} r^k P_x\{\tau_{K_t} > k\} \\ &= \sum_{k=1}^{\infty} r^k \sum_{i=k+1}^{\infty} P_x\{\tau_{K_t} = i\} \\ &= \sum_{i=2}^{\infty} P_x\{\tau_{K_t} = i\}\left(\sum_{k=1}^{i-1} r^k\right) \\ &= \frac{1}{r-1}\sum_{i=2}^{\infty} P_x\{\tau_{K_t} = i\}r^i - \frac{r}{r-1}P_x\{\tau_{K_t} > 1\} \\ &\geqq \frac{1}{r-1}(E_x[r^{\tau_{K_t}}] - rP_x\{\tau_{K_t} \leqq 1\}) - \frac{r}{r-1}P_x\{\tau_{K_t} > 1\} \\ &\geqq \frac{1}{r-1}E_x[r^{\tau_{K_t}}] - \frac{r}{r-1}.\end{aligned}$$

This together with (18) and the definitions of U and ε proves the lemma. □

Proof of Proposition 3.1. The existence of a moment satisfying (12) implies that

$$\limsup_{k\to\infty} E_x[V(\Phi_k)] \leqq \frac{K}{1-\lambda}$$

for all initial conditions, and hence also the stability condition (S2). By Proposition 2.2, $\mathbf{\Phi}$ is positive Harris recurrent.

Since V is a moment and is continuous, for each $t \in \mathbb{R}_+$, the set K_t defined above Lemma 3.1 is compact. By Lemma 2.2 we may choose t so large that K_t is petite, and so the proof follows from Proposition 2.1. □

In the following result we show that condition (12) is an extremely useful property even when the irreducibility condition is not satisfied. Proposition 3.2 will be used in §§ 4 and 5 to establish ergodic theorems for Markov chains that do not satisfy any recurrence condition.

PROPOSITION 3.2. *Suppose that V is a positive measurable function satisfying condition* (12). *Then the following inequalities are satisfied for every initial condition $x \in \mathbf{X}$:*

(i) $$P^k V(x) \leqq \lambda^k V(x) + \frac{K}{1-\lambda}, \qquad k \in \mathbb{Z}_+.$$

(ii) $$\limsup_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \sqrt{V(\Phi_k)} \leqq \frac{\sqrt{K}}{1-\sqrt{\lambda}} \quad a.s.\,[P_x].$$

Proof. The result (i) follows immediately from (12) and induction.

To show that (ii) holds, fix $x \in \mathbf{X}$, and let $U = \sqrt{V}$, $\alpha = \sqrt{\lambda}$, and $L = \sqrt{K}$. The function U is a moment and by Jensen's inequality,

$$\begin{aligned} PU(\Phi_k) &= E[U(\Phi_{k+1}) \mid \Phi_k] \quad \text{a.s.}\,[P_x] \\ &\leqq \sqrt{E[V(\Phi_{k+1}) \mid \Phi_k]} \\ &\leqq \sqrt{\lambda V(\Phi_k) + K} \\ &\leqq \alpha U(\Phi_k) + L. \end{aligned} \tag{19}$$

It follows that for every $N \in \mathbb{Z}_+$,

$$\frac{1}{N}\sum_{k=1}^{N} U(\Phi_k) \leqq \frac{1}{N}\sum_{k=1}^{N} (U(\Phi_k) - PU(\Phi_{k-1})) + \frac{1}{N}\sum_{k=1}^{N} (PU(\Phi_{k-1}) - PU(\Phi_k)) + \frac{1}{N}\sum_{k=1}^{N} \alpha U(\Phi_k) + L.$$

Let $\eta_k = U(\Phi_k) - PU(\Phi_{k-1})$, and observe that the second summand in the equation above is a telescoping series. Hence,

$$(1-\alpha)\frac{1}{N}\sum_{k=1}^{N} U(\Phi_k) \leqq \left|\frac{1}{N}\sum_{k=1}^{N} \eta_k\right| + \frac{1}{N}|PU(\Phi_0) - PU(\Phi_N)| + L. \tag{20}$$

It follows by result (i) of the theorem and a straight forward calculation that

$$\sum_{N=1}^{\infty} E_x\left[\frac{1}{N^2}|PU(\Phi_0) - PU(\Phi_N)|^2\right] \leqq 2\left(V(x) + \frac{K}{1-\lambda}\right)^2 \sum_{N=1}^{\infty} \frac{1}{N^2} < \infty,$$

and hence by Chebyshev's inequality and the Borel-Cantelli Lemma the second summand in the right-hand side of (20) converges to zero as $N \to \infty$ a.s. $[P_x]$.

Another straightforward calculation yields

$$\sup_{k \in \mathbb{Z}_+} E_x[\eta_k^2] \leqq 2 \sup_{k \in \mathbb{Z}_+} E_x[V(\Phi_k)] \leqq 2V(x) + \frac{2K}{1-\lambda}$$

by inequality (i) of the proposition. Hence with $\mathcal{F}_k$ defined in (13), the sequence $\{(\eta_k, \mathcal{F}_k): k \in \mathbb{Z}_+\}$ is an L^2 bounded martingale difference process. Applying Theorem 5.2 of Chapter 4 of Doob (1953), it follows that the first summand in the right-hand side of (20) converges to zero as $N \to \infty$. Taking "lim sup's" on either side of (20) yields (ii) of the theorem. □

We now show how a moment satisfying (12) may be constructed. In the stable linear case (5) with $\mathbf{w} = 0$ such a construction was first carried out in Kalman and Bertram (1960). Corollary 3.2* of that paper implies that there exists a positive definite matrix M with $I \leqq M \leqq mI$ for some $m \geqq 1$, and

$$|Ax|_M^2 \leqq \lambda |x|_M^2 \tag{21}$$

where $|y|_M^2 \triangleq y^T M y$ for $y \in \mathbb{R}^n$, and $\lambda < 1$. In fact, we can take

$$M \triangleq I + \sum_{i=1}^{\infty} A^{Ti} A^i.$$

In the case where $\mathbf{w} \neq 0$, suppose that the i.i.d. process $\mathbf{w}$ defined in (5) satisfies $E[w_0^2] < \infty$, but is otherwise arbitrary, and let V be the moment on $\mathbb{R}^n$ defined by

$$V(x) \triangleq x^T M x, \qquad x \in \mathbf{X}.$$

Then for each $x \in \mathbf{X}$ we have

$$PV(x) = x^T A^T M A x + E[w_0^T B^T M B w_0] \leqq \lambda V(x) + m|B|^2 E[w_0^2],$$

showing that the function V satisfies (12). We remark that in general, no stochastic Lyapunov function satisfying (14) exists in the linear case.

In the nonlinear case

$$\Phi_{k+1} = F(\Phi_k, w_{k+1})$$

with $\mathbf{X}=\mathbb{R}^n$ and $F:\mathbf{X}\times\mathbb{R}^p\to\mathbf{X}$ sufficiently smooth, it is possible to generalize this construction (see Chan (1986) for the details). However this requires extremely restrictive stability conditions. The idea is to look at the *freely evolving system*

$$d_{k+1}=F^{k+1}(x)\triangleq F(d_k,0),\qquad k\in\mathbb{Z}_+$$

with $d_0=x\in\mathbf{X}$ given. If there exists a fixed $0<\lambda<1$ and $K>0$ such that

$$|d_k|\leqq K\lambda^k|x|$$

for all $k\in\mathbb{Z}_+$ and $x\in\mathbf{X}$, then for fixed $\alpha\in(\lambda,1)$ we may define the moment V by

$$V(x)\triangleq\sup_{k\geqq 0}\alpha^{-k}|F^k(x)|$$

where $F^0(x)\triangleq x$. Under the appropriate conditions, V will satisfy condition (12).

Among the difficulties with this approach is that to apply the Lyapunov function for the freely evolving system to the original system, a global Lipschitz condition on V is needed. Presently, the only way to obtain such a condition is by imposing a global Lipschitz condition on F, and this is ruled out in many examples.

However, in all of the examples previously studied (see Guo and Meyn (1988) and the examples in § 5) a function satisfying (12) may be shown to exist by a reasonably simple calculation even though no global Lipschitz condition is satisfied.

In the next section we drop the recurrence condition, and show that suitably strong stability conditions (implied by the existence of a continuous moment V satisfying (12)) allow us to establish useful ergodic theorems for the class of continuous and bounded functions on $\mathbf{X}$.

4. Law of large numbers. Our main objective in this section is to find conditions under which the law of large numbers (2) holds for all initial conditions.

To illustrate the difficulties in finding such conditions when $\mathbf{\Phi}$ is not positive Harris recurrent, suppose that an invariant probability π exists. By a theorem of Doob (1953), for each $f\in L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\pi)$ there exists a function $f_\infty\in L^1(\mathbf{X},\mathcal{B}(\mathbf{X}),\pi)$ and a Borel set $\mathbf{X}_f\subset\mathbf{X}$ of full π measure such that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}f(\Phi_k)=f_\infty(x)\quad\text{a.s.}\,[P_x] \tag{22}$$

whenever $x\in\mathbf{X}_f$. The function f_∞ is a version of the conditional expectation $E[f|\Sigma_I]$ where Σ_I is the σ-algebra of π-invariant events in $\mathcal{B}(\mathbf{X})$. Define the *occupation probabilities*

$$\tilde{\mu}_N\{A\}\triangleq\frac{1}{N}\sum_{k=1}^{N}\mathbf{1}_{\Phi_k\in A},\quad N\in\mathbb{Z}_+,\quad A\in\mathcal{B}(\mathbf{X}), \tag{23}$$

and let $\{g_n:n\in\mathbb{Z}_+\}$ denote the bounded continuous functions defined in (7). It follows from (22) that there exists a Borel set $\mathbf{X}_1\subset\mathbf{X}$, and bounded Borel functions $\{g_{n,\infty}:n\in\mathbb{Z}_+\}$ such that whenever $x\in\mathbf{X}_1$,

$$\lim_{k\to\infty}\int g_n\,d\tilde{\mu}_k=g_{n,\infty}(x)\quad\text{a.s.}\,[P_x]$$

for each $n\in\mathbb{Z}_+$.

It also follows from (22) that for a.e. $[\pi]$ $x\in\mathbf{X}$ the sequence of probabilities $\{1/N\sum_{k=1}^{N}P^k(x,\cdot):N\in\mathbb{Z}_+\}$ is tight, and hence we may assume that these probabilities are tight for each $x\in\mathbf{X}_1$. By the Dominated Convergence Theorem we have

for each $x \in \mathbf{X}_1$ and $n \in \mathbb{Z}_+$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \int P^k(x, dy) g_n(y) = g_{n,\infty}(x),$$

and taking any weak limit point of the probabilities

$$\left\{ \frac{1}{N} \sum_{k=1}^{N} P^k(x, \cdot) : N \in \mathbb{Z}_+ \right\}, \tag{24}$$

it follows from this and (7) that for every $x \in \mathbf{X}_1$ there exists a probability π_x such that $g_{n,\infty}(x) = \int g_n \, d\pi_x$. Furthermore, for each $x \in \mathbf{X}_1$, every weak limit point of (24) is necessarily invariant (see, for example, the proof of Lemma 4.1) and hence π_x is an invariant probability for such x.

The following result follows from these observations.

PROPOSITION 4.1. *If the Feller Markov chain Φ possesses an invariant probability π, then there exists a Borel set $\mathbf{X}_1 \subset \mathbf{X}$ and invariant probabilities $\{\pi_x : x \in \mathbf{X}_1\}$ such that $\pi\{\mathbf{X}_1\} = 1$. The invariant probabilities $\{\pi_x\}$ have the property that for every initial condition $\Phi_0 = x \in \mathbf{X}_1$, the occupation probabilities converge*:

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi_x \quad \text{as } k \to \infty \quad \text{a.s. } [P_x].$$

Observe that the limit on the right-hand side of (22) is nonrandom, and hence so are the probabilities $\{\pi_x : x \in \mathbf{X}_1\}$. This result cannot be expected to hold for all initial conditions in general. For example, when Doeblin's condition (Condition D of Doob (1953)) or a stochastic controllability condition is satisfied [Meyn and Caines (1988)], it may be shown that there exists, at most, a countable collection of disjoint absorbing sets $\{H_i : i \in \mathbb{Z}_+\}$ on which the process Φ is positive Harris recurrent. For every initial condition $x \in \mathbf{X}$,

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \tilde{\pi}_x \quad \text{as } k \to \infty, \tag{25}$$

where the invariant probability $\tilde{\pi}_x$ is defined for $f \in \mathbf{C}$ by

$$\int f \, d\tilde{\pi}_x \triangleq \sum_{k=0}^{\infty} \left(\mathbf{1}_{\{\Phi \text{ enters } H_k\}} \int f \, d\pi_k \right), \tag{26}$$

and for each $i \in \mathbb{Z}_+$ the invariant probability π_i is supported on H_i. When $x \in H_i$ it may be shown that $\tilde{\pi}_x = \pi_i$, but, in general, $\tilde{\pi}_x$ will be a random probability.

In general, one of the main difficulties in establishing (25) is finding a candidate limit probability $\tilde{\pi}_x$. To this end suppose that an invariant Markov transition function Π exists (Π is a Markov transition function and $P\Pi = P\Pi = \Pi$) satisfying

$$\frac{1}{N} \sum_{k=1}^{N} P^k(x, \cdot) \xrightarrow{\text{weakly}} \Pi(x, \cdot) \quad \text{as } N \to \infty, \quad x \in \mathbf{X}. \tag{27}$$

Then with $\mathcal{F}_k$ defined in (13), the pair $(\Pi(\Phi_k, A), \mathcal{F}_k)$ is a convergent martingale, and we then make the definition

$$\tilde{\pi}_x\{A\} \triangleq \lim_{k\to\infty} \Pi(\Phi_k, A) \tag{28}$$

when $\Phi_0 = x \in \mathbf{X}$. For example, in the special case where Doeblin's condition holds we have $\Pi(\Phi_k, A) \to \pi_i\{A\}$ on the event $\{\Phi \text{ enters } H_i\}$, showing that this definition agrees with (26).

An interesting problem remains open. Suppose that Φ is a Feller Markov chain satisfying condition (S2). (i) Does an invariant Markov transition function satisfying (27) exist? And if so, (ii) does the law of large numbers (25) hold with $\tilde{\pi}_x$ defined in (28) for every initial condition?

We do not see any way of answering this question, and so for the time being we will have to settle for less general results.

At present, there are at least two possible routes to establishing the law of large numbers for continuous bounded functions. One is to assume that the Markov chain is *regular* in the sense of Feller (1971). This together with condition (S2) implies that an invariant Markov transition function satisfying (27) exists, and under general conditions (ii) follows using the same argument to be presented below. Alternatively, we may assume that a unique invariant probability π exists since if this is the case, then the problem of finding a candidate limit probability in (25) is solved.

Below we present a result based on the second approach since the first requires the introduction of additional definitions and preliminary results. The proof of the law of large numbers for regular Markov chains will appear in a planned monograph concerning Markov chains on topological spaces.

Below we state some assumptions that will be needed.

(A1) A unique invariant probability π exists.

(A2) The collection of probabilities on $\mathcal{B}(\mathbf{X})$,

$$\left\{\frac{1}{N}\sum_{k=1}^{N} P^k(x,\cdot): x\in K, N\in\mathbb{Z}_+\right\}$$

is tight for every compact set $K\subset\mathbf{X}$.

(A3) For all initial conditions $x\in\mathbf{X}$ there exists a sequence of compact sets $K_n\uparrow\mathbf{X}$ such that

$$\lim_{n\to\infty}\liminf_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\mathbf{1}_{\Phi_k\in K_n}=1 \quad \text{a.s.}\,[P_x].$$

Assumption (A3) is the condition that the occupation probabilities are almost surely tight for each initial condition. This condition will hold if a moment V exists for which

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}V(\Phi_k)<\infty \quad \text{a.s.}\,[P_x]$$

for each initial condition $x\in\mathbf{X}$. Assumption (A2) will be satisfied if a moment V exists with the property

$$\sup_{\substack{k\geqq 0\\ x\in K}} E_x[V(\Phi_k)]<\infty$$

for every compact set $K\subset\mathbf{X}$.

Hence applying Proposition 3.2, (A2) and (A3) will hold if a continuous moment V satisfying (12) exists.

We may now state the main result of this section that is a generalization of the result [Breiman (1960)] to noncompact state spaces.

PROPOSITION 4.2. *Suppose that conditions* (A1)-(A3) *hold. Then for each initial condition* $x\in\mathbf{X}$,

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi \quad \text{as } k\to\infty \quad \text{a.s.}\,[P_x].$$

We remark that in the case where f is continuous but unbounded it is often still possible to establish (2) under the conditions of Proposition 4.2. A sufficient condition is for some $\delta > 0$,

$$\limsup_{N\to\infty} \int |f|^{1+\delta}\, d\tilde{\mu}_N \triangleq \limsup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} |f(\Phi_k)|^{1+\delta} < \infty,$$

since this implies that the function f is uniformly integrable with respect to the occupation probabilities $\{\tilde{\mu}_k : k \in \mathbb{Z}_+\}$ with probability one (see Theorem 5.4 of Billingsley (1968)).

Under the conditions of Proposition 3.2,

$$\int \sqrt{V}\, d\pi \leqq \frac{\sqrt{L}}{1-\sqrt{\lambda}},$$

when π is an invariant probability. This implies that if V is a moment, the set of all invariant probabilities is tight. If, in addition, a uniquely invariant probability π exists, then by Proposition 4.2, $\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi$ with probability one. A slight modification of the proof of Proposition 3.2 shows that

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} V^p(\Phi_k) \leqq \frac{K^p}{1-\lambda^p} < \infty$$

for every $p \in (0, 1)$, which by uniform integrability implies that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} V^p(\Phi_k) = \int V^p\, d\pi \quad \text{a.s. } [P_x].$$

To prove Proposition 4.2 we will apply the following lemma.

LEMMA 4.1. *If conditions* (A1) *and* (A2) *hold then for every* $f \in \mathbf{C}$,

$$\frac{1}{N} \sum_{k=1}^{N} P^k f(\cdot) \to \int f\, d\pi$$

as $N \to \infty$, *uniformly on compact sets.*

Proof of Lemma 4.1. Suppose that (A1) and (A2) hold, and for $f \in \mathbf{C}$, $\delta > 0$, and $N \in \mathbb{Z}_+$ define

$$A_N^{\delta}(f) \triangleq \left\{ x \in \mathbf{X} : \left| \frac{1}{N} \sum_{k=1}^{N} P^k f(x) - \int f\, d\pi \right| \geqq \delta \right\}.$$

The proof is by contradiction. If the conclusion of the lemma is false, then there exists f, δ, a compact set $K \subset \mathbf{X}$, and a subsequence $\{N_i\}$ of $\mathbb{Z}_+$ such that

$$A_{N_i}^{\delta}(f) \cap K \neq \phi$$

for all $i \in \mathbb{Z}_+$.

For each $i \in \mathbb{Z}_+$ let $x_i \in A_{N_i}^{\delta}(f) \cap K$, and consider the sequence of probabilities

$$\left\{ \frac{1}{N_i} \sum_{k=1}^{N_i} P^k(x_i, \cdot) : i \in \mathbb{Z}_+ \right\}.$$

This collection is tight by assumption, and hence by choosing a further subsequence if necessary we may assume that a probability λ exists for which

$$\frac{1}{N_i} \sum_{k=1}^{N_i} P^k(x_i, \cdot) \xrightarrow{\text{weakly}} \lambda \quad \text{as } i \to \infty.$$

We will now show that λ is an invariant probability. For each $g \in \mathbf{C}$ we have by weak convergence,

$$\int\int P(x, dy)g(y)\lambda(dx) = \lim_{i\to\infty} \frac{1}{N_i} \sum_{k=1}^{N_i} \int P^{k+1}(x_i, dy)g(y)$$

$$= \int g\, d\lambda.$$

Applying Urysohn's Lemma and using the fact that probabilities on $\mathcal{B}(\mathbf{X})$ are regular, it may be shown that characteristic functions of Borel sets may be approximated in $L^1(\lambda)$ and $L^1(\lambda P)$ by elements of $\mathbf{C}$. It follows that λ is invariant, and applying (A1) we conclude that $\lambda = \pi$. However by construction of λ, δ, and the function f,

$$\left| \int f\, d\pi - \int f\, d\lambda \right| = \lim_{i\to\infty} \left| \frac{1}{N_i} \sum_{k=1}^{N_i} \int P^k f\,(x_i) - \int f\, d\pi \right| \geqq \delta > 0.$$

This contradiction proves the lemma. □

Proof of Proposition 4.2. Fix $f \in \mathbf{C}$ and $\Phi_0 = x \in \mathbf{X}$. For each N, $n \in \mathbb{Z}_+$ we have

$$\frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f\, d\pi = \sum_{i=0}^{n-1} \frac{1}{N} \sum_{k=1}^{N} (P^i f(\Phi_{k-i}) - P^{i+1} f(\Phi_{k-i-1}))$$

$$+ \frac{1}{N} \sum_{k=1}^{N} P^n f(\Phi_k) - \int f\, d\pi$$

$$+ \frac{1}{N} \sum_{k=1}^{N} (P^n f(\Phi_{k-n}) - P^n f(\Phi_k))$$

where we adopt the convention that $\Phi_k = \Phi_0$ for $k \leqq 0$. For each $M \in \mathbb{Z}_+$ we may average the right-hand side of this equality from $n = 1$ to M to obtain

$$\frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f\, d\pi = \frac{1}{M} \sum_{n=1}^{M} \left(\sum_{i=0}^{n-1} \frac{1}{N} \sum_{k=1}^{N} (P^i f(\Phi_{k-i}) - P^{i+1} f(\Phi_{k-i-1})) \right)$$

$$+ \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{M} \sum_{n=1}^{M} P^n f(\Phi_k) \right) - \int f\, d\pi$$

$$+ \frac{1}{M} \sum_{n=1}^{M} \left(\frac{1}{N} \sum_{k=1}^{N} P^n f(\Phi_{k-n}) - P^n f(\Phi_k) \right).$$

The third term is a telescoping series, and hence letting $\|f\|_\infty \triangleq \sup_{x\in\mathbf{X}} |f(x)|$ we have

$$\left| \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f\, d\pi \right| \leqq \sum_{i=0}^{M-1} \left| \frac{1}{N} \sum_{k=1}^{N} (P^i f(\Phi_{k-i}) - P^{i+1} f(\Phi_{k-i-1})) \right|$$

$$+ \left| \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{M} \sum_{n=1}^{M} P^n f(\Phi_k) \right) - \int f\, d\pi \right| + \frac{2M\|f\|_\infty}{N}. \tag{29}$$

For each fixed $0 \leqq i \leqq M-1$ the sequence

$$(P^i f(\Phi_{k-i}) - P^{i+1} f(\Phi_{k-i-1}), \mathcal{F}_{k-i}), \qquad k > i,$$

is a bounded martingale difference process, where $\mathcal{F}_k$ is defined in (13). Hence by Theorem 5.2 of Doob (1953, Chap. 4), the first summand converges to zero almost surely for every $M \in \mathbb{Z}_+$. Let $\varepsilon > 0$, and $\{\Phi_k\}$ be a sample for which the first summand

converges to zero for each M. By (A3), for almost every sample path there exists a compact set $K \subset \mathbf{X}$ such that

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{\Phi_k \in K\}} \geqq 1 - \varepsilon.$$

Fix such a set K, and choose M so large that $|1/M \sum_{n=1}^{M} P^n f(x) - \int f\, d\pi| < \varepsilon$ on K. This is possible by (A1) and (A2), and Lemma 4.1. Hence by (29),

$$\begin{aligned}
&\limsup_{N\to\infty} \left| \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f\, d\pi \right| \\
&\quad \leqq \limsup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \left| \left(\frac{1}{M} \sum_{n=1}^{M} P^n f(\Phi_k) \right) - \int f\, d\pi \right| \\
&\quad \leqq \limsup_{N\to\infty} \|f\|_\infty \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{\Phi_k \in K^c\}} \\
&\qquad + \limsup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{\Phi_k \in K\}} \left| \left(\frac{1}{M} \sum_{n=1}^{M} P^n f(\Phi_k) \right) - \int f\, d\pi \right| \\
&\quad \leqq \varepsilon \|f\|_\infty + \varepsilon,
\end{aligned}$$

and this shows that (2) holds for all $f \in \mathbf{C}$ and all $x \in \mathbf{X}$.

Let $\{g_n\}$ be the continuous functions defined in (7). Then since (2) holds for each bounded and continuous function,

$$P_x \left\{ \lim_{k\to\infty} \int g_n\, d\tilde{\mu}_k = \int g_n\, d\pi \quad \text{for each } n \in \mathbb{Z}_+ \right\} = 1.$$

This together with (7) proves the theorem. □

We now give alternative sufficient conditions to establish (2) for continuous and bounded functions. Let $d(\cdot, \cdot)$ denote the metric on $\mathbf{X}$, and define

$$d(x, E) \triangleq \inf_{y \in E} d(x, y) \quad \text{and} \quad B_\delta(E) \triangleq \{y \in \mathbf{X} : d(y, E) < \delta\},$$

for $x \in \mathbf{X}$, $E \in \mathcal{B}(\mathbf{X})$, and $\delta > 0$.

In cases where a unique invariant probability π exists, the following conditions are often satisfied.

(A4) There exists a closed set $H \subset \mathbf{X}$ that supports π, and H is a Harris set. Furthermore, the support of π has nonempty interior in H.

(A5) For all initial conditions $\Phi_0 = x \in \mathbf{X}$,

$$\lim_{k\to\infty} d(\Phi_k, H) = 0 \quad \text{a.s. } [P_x].$$

PROPOSITION 4.3. *Suppose that conditions* (A1), (S2) *and* (A3)-(A5) *hold, and that the Markov chain* Φ *restricted to* H *has period* $r \geqq 1$. *Then for each initial condition* $x \in \mathbf{X}$,

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi \quad a.s.\ [P_x],$$

$$\frac{1}{r} \sum_{i=1}^{r} P^{k+i}(x, \cdot) \xrightarrow{\text{weakly}} \pi \quad as\ k \to \infty.$$

The idea of the proof is basically the same as that of Proposition 4.2. We show that for any compact set $K \subset \mathbf{X}$ the probability $1/r \sum_{i=1}^{r} P^{k+i}(x, \cdot)$ is close to π in the weak topology for all x in a neighborhood of $K \cap H$, and all $k \in \mathbb{Z}_+$ sufficiently large. The proof then follows using the same martingale difference argument as in Proposition 4.2.

The following result follows from Corollary 4.1 and Theorem 2.1 of Cogburn (1975).

LEMMA 4.2. *If condition* (A4) *holds, then for every* $\varepsilon > 0$ *and every compact set* $K \subset \mathbf{X}$, *there exists* $k = k(K) \in \mathbb{Z}_+$ *such that*

$$\left\| \frac{1}{r} \sum_{i=1}^{r} P^{k+i}(x, \cdot) - \pi(\cdot) \right| < \varepsilon$$

for every $x \in K \cap H$.

Proof of Proposition 4.3. Let $\varepsilon > 0$, let $K \subset \mathbf{X}$ be a compact set, and fix $f \in \mathbf{C}$. Under (A4) we may apply Lemma 4.2 to find $k_0 \in \mathbb{Z}_+$ such that

$$\left| \frac{1}{r} \sum_{i=1}^{r} P^{k_0+i} f(x) - \int f \, d\pi \right| < \varepsilon \tag{30}$$

for all $x \in H \cap K$, and by the Feller property it follows that for some $\delta > 0$, (30) holds for all $x \in B_\delta(H) \cap K$.

For all $n \in \mathbb{Z}_+$,

$$\frac{1}{r} \sum_{i=1}^{r} P^{k_0+i+n} f(x) = P^n \left(\frac{1}{r} \sum_{i=1}^{r} P^{k_0+i} f \right)(x).$$

Hence by (30) for every $x \in \mathbf{X}$,

$$\begin{aligned} &\limsup_{n \to \infty} \left| \frac{1}{r} \sum_{i=1}^{r} P^{k_0+i+n} f(x) - \int f \, d\pi \right| \\ &\quad \leqq \limsup_{n \to \infty} \varepsilon P^n(x, K \cap B_\delta(H)) + \|f\|_\infty (P^n(x, B_\delta^c(H)) + P^n(x, K^c)). \end{aligned}$$

By (S2) and (A5), and since $\varepsilon > 0$ and K compact are arbitrary, this proves the first assertion of the proposition.

By (29), (30), and applying the martingale difference argument used in the proof of Proposition 4.2, we have

$$\begin{aligned} &\limsup_{N \to \infty} \left| \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f \, d\pi \right| \\ &\quad \leqq \varepsilon + \|f\|_\infty \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\Phi_k \notin K \cap B_\delta(H)} \\ &\quad \leqq \varepsilon + \|f\|_\infty \left(\limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\Phi_k \notin K} + \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\Phi_k \notin B_\delta(H)} \right). \end{aligned}$$

Since we have assumed that $\mathbf{\Phi}$ converges to the set H with probability one, it follows that

$$\limsup_{N \to \infty} \left| \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) - \int f \, d\pi \right| \leqq \|f\|_\infty \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\Phi_k \notin K} + \varepsilon.$$

By (A3), and since $\varepsilon > 0$ is arbitrary, it follows that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) = \int f d\pi \quad \text{a.s. } [P_x]$$

for each $f \in \mathbf{C}$ and, as was shown in the proof of Proposition 4.2, this implies that

$$\bar{\mu}_k \xrightarrow{\text{weakly}} \pi \quad \text{a.s. } [P_x]. \qquad \square$$

5. Examples. In this section we consider a number of examples to illustrate how the results above may be applied to specific control problems. With the exception of the example presented in § 5.2 where slightly different techniques are used, we assume that the Markov chain $\mathbf{\Phi}$ has the form

$$\Phi_{k+1} = F(\Phi_k, w_{k+1}), \qquad k \in \mathbb{Z}_+. \tag{1}$$

For all k, $\Phi_k \in \mathbf{X}$, a d-dimensional manifold, $w_k \in \mathbb{R}^p$, and $F : \mathbf{X} \times \mathbb{R}^p \to \mathbf{X}$ is continuously differentiable (C^1).

We further assume that $(\Phi_0, \mathbf{w})$ are random variables on the probability space $(\Omega, \mathcal{F}, P_{\Phi_0})$, Φ_0 is independent of $\mathbf{w}$, and that $\mathbf{w}$ is an independent and identically distributed process.

We will also require the following conditions on the distribution μ_w of the random variables w_k, $k \in \mathbb{Z}_+$, whenever the weak stochastic controllability condition is required:

(W1) The distribution μ_w of w_k, $k \in \mathbb{Z}_+$, possesses a density that is lower semicontinuous;

(W2) $0 \in \text{supp } \mu_w$;

Condition (W1) implies that μ_w possesses a density that is strictly positive on an open set $\mathcal{O}_w \subset \mathbb{R}^p$ and zero elsewhere, and hence supp $\mu_w = \bar{\mathcal{O}}_w$.

Before presenting the examples, we give some useful definitions from Meyn and Caines (1988), and present a result from that paper that will be needed in the examples that follow.

Here we give a precise definition of weak stochastic controllability. General conditions for weak stochastic controllability involving a controllability matrix may be found in Meyn and Caines (1988). Given two measures ν and μ on $\mathcal{B}(\mathbf{X})$ we say that ν is *absolutely continuous* with respect to μ (denoted $\nu \prec \mu$) if $\nu\{A\} = 0$ whenever $\mu\{A\} = 0$. We let $\mathbf{1}_A \mu$ denote the measure defined for $B \in \mathcal{B}(\mathbf{X})$ by $(\mathbf{1}_A \mu)\{B\} \triangleq \mu\{A \cap B\}$.

DEFINITION. The system (1) is called *weakly stochastically controllable* if for each initial condition $x \in \mathbf{X}$ there exists $T = T(x) \in \mathbb{Z}_+$, and an open set $\mathcal{O}_x \subset \mathbf{X}$ such that $P^T(x, \cdot) > \mathbf{1}_{\mathcal{O}_x} \mu^{Leb}$.

Hence if $\mathbf{\Phi}$ is weakly stochastically controllable, then the Radon-Nikodym derivative of the probability $P^T(x, \cdot)$ (with respect to Lebesgue measure) is strictly positive on an open set $\mathcal{O}_x \subset \mathbf{X}$.

We call the deterministic system

$$d_{k+1} = F(d_k, 0), \qquad k \in \mathbb{Z}_+ \tag{31}$$

with initial condition $d_0 \in \mathbf{X}$ the *freely evolving system.* The system (1) *satisfies condition* (GA) if there exists $d^* \in \mathbf{X}$ that is globally attracting for the freely evolving system. That is,

(GA) For each initial condition $x \in \mathbf{X}$, $\lim_{k\to\infty} d_k = d^*$.

Hence, if the disturbance sequence $\mathbf{w}$ is replaced by $(0, \cdots, 0, \cdots)$ in (1) then $\Phi_k \to d^*$ as $k \to \infty$ for all initial conditions.

The following result will be used together with Proposition 3.1 in the examples below.

PROPOSITION 5.1. *Suppose that $\mathbf{\Phi}$ is a Markov chain of the form* (1), *and that conditions* (W1) *and* (W2) *hold. If $\mathbf{\Phi}$ is weakly stochastically controllable and satisfies condition* (GA), *then the irreducibility condition is satisfied for an open set A containing d^*, and $\mathbf{\Phi}$ is aperiodic.*

We may now proceed with the first example.

5.1. Nonlinear control. Here we consider a linear single-input single-output stochastic state space system with nonlinear feedback control law

$$u_k \triangleq -\varphi(y_k) \quad \text{for all } k \in \mathbb{Z}_+ \tag{32}$$

where the function $\varphi \in C^1$. We assume that $\varphi(0) = 0$, and to simplify the analysis we also take $d\varphi/dt(0) \neq 0$.

The closed loop system equations are

$$\begin{aligned} x_{k+1} &= Ax_k - b\varphi(c^{\mathrm{T}}x_k + \zeta_{k+1}) + G\xi_{k+1}, \\ y_k &= c^{\mathrm{T}}x_k + \zeta_{k+1}, \qquad k \in \mathbb{Z}_+, \end{aligned} \tag{33}$$

and it is easily seen that if $\mathbf{w} \triangleq \binom{\xi}{\zeta}$ satisfies the conditions given at the beginning of this section, then $\mathbf{x}$ is a Feller Markov chain of the form (1) with state space $\mathbb{R}^n$.

In fact, $\mathbf{\Phi} \triangleq \binom{\mathbf{x}}{\mathbf{y}}$ will also be a Markov chain under the appropriate conditions whose state space $\mathbf{X} \triangleq \mathbb{R}^{n+1}$. However, we may show that almost any result of interest obtainable for the process $\mathbf{x}$ will carry over to the joint process $\mathbf{\Phi}$, and so we restrict our analysis to the simpler Markov chain.

This is a popular example in nonlinear systems theory (see, e.g., Zames (1966), Popov (1973), and Safonov (1980)) and is ideal for illustrating the general results presented in the previous sections.

The following stability and controllability conditions will be needed below. We say that the control φ defined in (32) lies in the sector (α, r) (see Safonov (1980)) if for all $x \in \mathbb{R}$,

$$|\varphi(x) - \alpha x| \leqq r|x|.$$

For a positive definite $n \times n$ matrix Q, a vector $z \in \mathbb{R}^n$, and an $n \times n$ matrix F we let

$$|z|_Q^2 = z^{\mathrm{T}}Qz \quad \text{and} \quad |F|_Q^2 = \sup_{z \neq 0} \frac{|Fz|_Q^2}{|z|_Q^2}.$$

(NC1) $E[|w_0|^{2+\delta}] < \infty$ for some $\delta > 0$:

(NC2) The control law φ lies in the sector (α, r), and for some positive definite $n \times n$ matrix Q,

$$\lambda \triangleq |(A - \alpha bc^{\mathrm{T}})|_Q + r|b|_Q|c|_{Q^{-1}} < 1.$$

(NC3) The pair $(A, [G|b])$ is controllable;

(NC4) The distribution μ_w of w_0 satisfies conditions (W1) and (W2).

Let P denote the Markov transition function on $\mathbf{X} = \mathbb{R}^{n+1}$ for the joint process $\mathbf{\Phi}$, and let $\tilde{\mu}_k$, $k \in \mathbb{Z}_+$, denote the occupation probabilities defined in § 4. The functions x, u, and y on $\mathbf{X}$ are defined so that

$$x_k = x(\Phi_k), \quad u_k = u(\Phi_k), \quad y_k = y(\Phi_k), \quad k \in \mathbb{Z}_+.$$

Our objective in this section is to prove Proposition 5.2.

PROPOSITION 5.2. *Suppose that conditions* (NC1), (NC2), *and* (NC4) *hold for* (33). *Then a unique invariant probability* π *exists, and the following limits hold for each initial condition* $x \in \mathbb{R}^{n+1}$

$$P^k(x, \cdot) \xrightarrow{\text{weakly}} \pi, \tag{34}$$

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi \quad a.s.\,[P_x], \tag{35}$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} (x_k^2 + y_k^2 + u_k^2) = \int (x^2 + y^2 + u^2)\, d\pi \quad a.s.\,[P_x], \tag{36}$$

$$\lim_{k\to\infty} E_x[x_k^2 + y_k^2 + u_k^2] = \int (x^2 + y^2 + u^2)\, d\pi. \tag{37}$$

If in addition (NC3) *holds, then* $\mathbf{\Phi}$ *is geometrically ergodic, and* (37) *holds at a geometric rate.*

Proposition 5.2 will be proved in several steps. We first present sufficient conditions for the Markov chain $\mathbf{x}$ to be weakly stochastically controllable.

5.1.1. Controllability. The generalized controllability matrix associated with $\mathbf{x}$ (see Meyn and Caines (1988) and Meyn (1987)) is defined for an initial condition $x \in \mathbb{R}^n$ by

$$C_x^T = [A_{T-1} \cdots A_1 B_0 | A_{T-1} \cdots A_2 B_1 | \cdots | A_{T-1} B_{T-2} | B_{T-1}] \tag{38}$$

where, letting $\alpha_k \triangleq d\varphi/dt\,(y_k)$,

$$\begin{aligned} A_k &\triangleq \left[\frac{\partial F}{\partial x}\right]_{(x_k, w_{k+1})} = A - \alpha_k b c^{\mathsf{T}} \\ B_k &\triangleq \left[\frac{\partial F}{\partial z}\right]_{(x_k, w_{k+1})} = [G | -\alpha_k b] \end{aligned} \qquad \text{for all } k \in \mathbb{Z}_+. \tag{39}$$

Observe that the generalized controllability matrix C_x^T is a function of the random variables $\{y_k : 0 \leqq k \leqq T-1\}$, and hence is also random. By Theorem 2.1 of Meyn and Caines (1988), $\mathbf{x}$ is weakly stochastically controllable if for each $x \in \mathbb{R}^n$ there exists $T \in \mathbb{Z}_+$ such that the matrix C_x^T has rank n with positive probability.

The following lemma greatly simplifies the computation of the rank of the matrix C_x^T. For an $n \times m$ matrix H let co-ker (H) denote the n-dimensional vector space

$$\text{co-ker}\,(H) \triangleq \{x \in \mathbb{R}^n : x^{\mathsf{T}} H = 0\}.$$

LEMMA 5.1. *The generalized controllability matrix* C_x^T *satisfies*

$$\text{co-ker}\,(C_x^T) = \text{co-ker}\,([A^{T-1}[G|\alpha_0 b]| \cdots |[A[G|\alpha_{T-2} b]|[G|\alpha_{T-1} b]]) \tag{40}$$

and hence $\mathbf{x}$ *is weakly stochastically controllable under conditions* (NC3) *and* (NC4) *if* $d\varphi/dt(0) \neq 0$.

Proof of Lemma 5.1. We will proceed by showing inductively that for $k = 0, \cdots, T-1$ and $x \in \mathbb{R}^n$,

$$\begin{aligned} &x^{\mathsf{T}}[A_{T-1}^x \cdots A_{T-k}^x B_{T-k-1}^x | \cdots | A_{T-1}^x B_{T-2}^x | B_{T-1}^x] = 0 \\ &\text{if and only if} \\ &x^{\mathsf{T}}[A^k[G|\alpha_{T-k-1} b]| \cdots |A[G|\alpha_{T-2} b]|[G|\alpha_{T-1} b]] = 0. \end{aligned} \tag{41}$$

For $k=0$, (41) becomes

$$x^{\mathrm{T}}[G\,|\,\alpha_{T-1}b]=0 \Leftrightarrow x^{\mathrm{T}}[G\,|\,-\alpha_{T-1}b]=0,$$

and this is obvious. Suppose now that (41) has been established for $k=n-1\geqq 0$. To establish the implication ($\Rightarrow$) for $k=n$ observe that if $x\in\mathbb{R}^n$ satisfies

$$x^{\mathrm{T}}[A^x_{T-1}\cdots A^x_{T-n}B^x_{T-n-1}|\cdots|A^x_{T-1}B^x_{T-2}\,|\,B^x_{T-1}]=0 \tag{42}$$

then by the induction hypothesis,

$$x^{\mathrm{T}}[A^{n-1}[G\,|\,\alpha_{T-n}b]|\cdots|A[G\,|\,\alpha_{T-2}b]|[G\,|\,\alpha_{T-1}b]]=0. \tag{43}$$

Furthermore, by (42) and (43) it follows that

$$\begin{aligned}0&=x^{\mathrm{T}}[A^x_{T-1}\cdots A^x_{T-n}B^x_{T-n-1}]\\&=x^{\mathrm{T}}(A-\alpha_{T-1}bc^{\mathrm{T}})(A-\alpha_{T-2}bc^{\mathrm{T}})\cdots(A-\alpha_{T-n}bc^{\mathrm{T}})[G\,|\,-\alpha_{T-1-n}b]\\&=x^{\mathrm{T}}A^n[G\,|\,-\alpha_{T-1-n}b].\end{aligned}$$

This and (43) establishes the implication ($\Rightarrow$) in (41) when $k=n$.

To establish the reverse implication suppose that $x\in\mathbb{R}^n$ satisfies

$$x^{\mathrm{T}}A^i[G\,|\,\alpha_{T-1-i}b]=0 \quad \text{for all } 0\leqq i\leqq n, \tag{44}$$

so that by the induction hypothesis

$$x^{\mathrm{T}}[A^x_{T-1}\cdots A^x_{T-n+1}B^x_{T-n}|\cdots|B^x_{T-1}]=0. \tag{45}$$

To complete the proof of the lemma we are left to show that

$$x^{\mathrm{T}}A^x_{T-1}\cdots A^x_{T-n}B^x_{T-n-1}=0,$$

and this follows from (39) and (44):

$$\begin{aligned}&x^{\mathrm{T}}A^x_{T-1}\cdots A^x_{T-n+1}B^x_{T-n}\\&\quad=x^{\mathrm{T}}(A-\alpha_{T-1}bc^{\mathrm{T}})(A-\alpha_{T-2}bc^{\mathrm{T}})\cdots(A-\alpha_{T-n}bc^{\mathrm{T}})[G\,|\,-\alpha_{T-1-n}b]\\&\quad=x^{\mathrm{T}}A^n[G\,|\,-\alpha_{T-1-n}b]=0.\end{aligned}$$

This establishes the first part of the lemma. If (NC3) and (NC4) hold, and if $d\varphi/dt(0)\neq 0$, then by (40) it follows that the matrix $C^n_0(0)$ has rank n. Here $C^n_0(0)$ denotes the generalized controllability matrix at time n evaluated along the output sequence $\mathbf{y}\equiv 0$. Hence by Corollary 4.4 of Meyn and Caines (1988), the Markov chain $\mathbf{x}$ is weakly stochastically controllable. □

5.1.2. Stability. We now show that a moment on $\mathbb{R}^n$ exists that satisfies (12). Let Q be the matrix defined in (NC2), and let $V(\cdot)\triangleq|\cdot|_Q$.

LEMMA 5.2. *Suppose that conditions* (NC1) *and* (NC2) *are satisfied. Then*

(i) *Condition* (GA) *holds with* $d^*=0$.

(ii) *The moment* V *satisfies* (12).

(iii) *For all initial conditions*

$$\mathbf{\Phi}_0=\begin{pmatrix}x_0\\y_0\end{pmatrix}\in\mathbf{X}:$$

$$\sup_{k\in\mathbb{Z}_+}E_{\mathbf{\Phi}0}[|x_k|^{2+\delta}+|u_k|^{2+\delta}+|y_k|^{2+\delta}]<\infty,$$

$$\limsup_{N\to\infty}\int|x|^{2+\delta}+|u|^{2+\delta}+|y|^{2+\delta}\,d\tilde{\mu}_N$$

$$\triangleq\limsup_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}(|x_k|^{2+\delta}+|u_k|^{2+\delta}+|y_k|^{2+\delta})<\infty \quad a.s.\,[P_{\Phi_0}]$$

where $\delta>0$ *is the constant used in condition* (NC1).

Proof. Equation (33) and (NC2) imply that

$$
\begin{aligned}
|x_{k+1}|_Q &\leqq |A-\alpha bc^{\mathsf{T}}|_Q|x_k|_Q + r|b|_Q|c^{\mathsf{T}}x_k+\zeta_{k+1}| + |\alpha||b|_Q|\zeta_{k+1}| + |G|_Q|\xi_{k+1}|_Q \\
&\leqq |A-\alpha bc^{\mathsf{T}}|_Q|x|_Q + r|b|_Q(|c|_{Q^{-1}}|x_k|_Q + |\zeta_{k+1}|) \\
&\qquad + |\alpha||b|_Q|\zeta_{k+1}| + |G|_Q|\xi_{k+1}|_Q \\
&\leqq \lambda|x_k|_Q + (|\alpha|+r)|b|_Q|\zeta_{k+1}| + |G|_Q|\xi_{k+1}|_Q.
\end{aligned} \tag{46}
$$

It immediately follows that (GA) holds, and that

$$
PV(x) \leqq \lambda V(x) + E[(\alpha+r)|b|_Q|\zeta_1| + |G|_Q|\xi_1|_Q]
$$

so that V satisfies (12).

Finally, if $\delta > 0$ is the constant used in (NC1), then (46) implies that for a constant B_1, and a random constant B_2,

$$
(E_x[|x_{k+1}|_Q^{2+\delta}])^{1/(2+\delta)} \leqq \lambda(E_x[|x_k|_Q^{2+\delta}])^{1/(2+\delta)} + B_1, \qquad k \in \mathbb{Z}_+,
$$

and

$$
\left(\frac{1}{N+1}\sum_{k=1}^{N+1}|x_k|_Q^{2+\delta}\right)^{1/(2+\delta)} \leqq \lambda\left(\frac{1}{N}\sum_{k=1}^{N}|x_k|_Q^{2+\delta}\right)^{1/(2+\delta)} + B_2, \qquad N \in \mathbb{Z}_+.
$$

By a geometric series argument, this shows that the third assertion of the lemma holds. □

We may now prove Proposition 5.2.

Proof of Proposition 5.2. We first suppose that (NC1)-(NC4) are satisfied. If this is the case then by Lemmas 5.1 and 5.2 and Proposition 5.1, the conditions of Proposition 3.1 are satisfied and hence $\mathbf{x}$ is geometrically ergodic. This implies that the joint process $\mathbf{\Phi} = \binom{\mathbf{x}}{\mathbf{y}}$ is also geometrically ergodic since $\mathbf{y}$ is virtually a function of $\mathbf{x}$.

Result (36) follows from this fact together with Proposition 1.1. To show that the convergence result (37) holds at a geometric rate, apply Proposition 2.1(iv).

We now relax condition (NC3). If the pair $(A, [G|b])$ is not controllable, then $\mathbf{x}$ may be decomposed into controllable and uncontrollable parts using a similarity transformation M where

$$
MAM^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},
$$

$$
M[G|b] = \begin{bmatrix} G_1 & b_1 \\ 0 & 0 \end{bmatrix},
$$

and $(A_{11}, [G_1|b_1])$ is controllable.

Letting

$$
\begin{pmatrix} x_k^c \\ x_k^0 \end{pmatrix} \triangleq Mx_k \quad \text{and} \quad c_1^{\mathsf{T}} = c^{\mathsf{T}}M^{-1}
$$

it follows that

$$
x_{k+1}^c = A_{11}x_k^c + A_{12}x_k^0 + b_1\varphi\left(c_1^{\mathsf{T}}\begin{pmatrix} x_k^c \\ x_k^0 \end{pmatrix} + \zeta_{k+1}\right) + G_1 w_{k+1},
$$

$$
x_{k+1}^0 = A_{22}x_k^0.
$$

If $x_0^0 = 0$ then $x_k^0 = 0$ for all $k \in \mathbb{Z}_+$, and in this case $\binom{\mathbf{x}^c}{\mathbf{y}}$ becomes a Markov chain for which the analysis above is valid.

By condition (GA), for all initial conditions $x \in \mathbf{X}$, $x_k^0 \to 0$ as $k \to 0$, and it follows that there exists a unique invariant probability π for $\mathbf{\Phi}$ under which $P_\pi\{\mathbf{x}^0 \equiv 0\} = 1$.

This shows that a hyperplane $H \subset \mathbb{R}^n$ exists that is a Harris set for the Markov chain $\mathbf{x}$. This set is necessarily a closed subset of $\mathbb{R}^n$, and by weak stochastic controllability (of $\mathbf{x}$ restricted to H) it follows that the support of π has nonempty interior in H. Hence (A4) and (A5) hold, and applying Proposition 4.3 completes the proof. □

5.2. Random linear system. Here we consider a randomly disturbed linear system

$$x_{k+1} = F(\xi_k)x_k + w_{k+1} \tag{47}$$

where $\mathbf{x}$ evolves on $\mathbb{R}^n$, ξ is a Feller Markov chain evolving on a compact m-dimensional manifold M, the disturbance process $\mathbf{w}$ is i.i.d., and $F: M \to gl(n, \mathbb{R})$ is continuous.

We also require these further assumptions:

(RLS1) $E[|w_1|^{2+\delta}] < \infty$ for some $\delta > 0$;
(RLS2) The processes $\mathbf{w}$ and ξ are independent;
(RLS3) ξ is aperiodic and positive Harris recurrent;
(RLS4) There exists $T_0 \in \mathbb{Z}_+$ and $\lambda_0 < 1$ such that
$E_{\xi_0}[F(\xi_0)^\tau \cdots F(\xi_{T_0})^\tau F(\xi_{T_0}) \cdots F[(\xi_0)] < \lambda_0 I$
for each $\xi_0 \in M$.

A similar example is treated in Feigin and Tweedie (1985), where geometric ergodicity is established under the condition that $\boldsymbol{\xi}$ is i.i.d., together with a nonsingularity assumption on the distribution of w_0.

It is easy to see that under the present conditions, $\mathbf{\Phi}$ is a Feller Markov chain with state space $\mathbf{X} \triangleq \mathbb{R}^n \times M$. Its Markov transition function P may be defined in terms of the Markov transition function Q for $\boldsymbol{\xi}$, and the system description (47).

Our main objective is to establish Proposition 5.3 below.

Let $x: \mathbf{X} \to \mathbb{R}^n$ denote the coordinate variable defined so that $x(\Phi_k) = x_k$, and let $\tilde{\mu}_k$, $k \in \mathbb{Z}_+$, denote the occupation probabilities defined in § 4.

PROPOSITION 5.3. *Suppose that the Markov chain* $\mathbf{\Phi} = \binom{x}{\xi}$ *defined in* (47) *satisfies conditions* (RLS1)-(RLS4). *Then there exists a unique invariant probability* π *on* $\mathbb{R}^n \times M$, *and for each initial condition* $x \in \mathbf{X}$,

$$P^k(x, \cdot) \xrightarrow{\text{weakly}} \pi \quad \text{as } k \to \infty, \tag{48}$$

$$\tilde{\mu}_k \xrightarrow{\text{weakly}} \pi \quad \text{as } k \to \infty \quad \text{a.s. } [P_x], \tag{49}$$

$$\lim_{k\to\infty} E_x[x_k^2] = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} x_k^2 = \int x^2 \, d\pi < \infty \quad \text{a.s. } [P_x]. \tag{50}$$

Observe that the statistical assumptions (RLS2) and (RLS3) and the stability assumptions (RLS1) and (RLS4) are all that is needed to establish the existence of a unique invariant probability π for $\mathbf{\Phi}$. The distributional assumptions (W1) and (W2) and the stochastic controllability hypothesis are not needed, and hence $\mathbf{\Phi}$ does not necessarily satisfy the irreducibility hypothesis.

We remark that under slightly stronger assumptions (conditions (W1) and (W2), and the condition that the support of the invariant probability for ξ has nonempty interior in M) the Markov chain $\mathbf{\Phi}$ will be geometrically ergodic. This follows from Lemma 5.3 below, and Proposition 3.1 applied to the sampled process $\{\Phi_{kT_0}: k \in \mathbb{Z}_+\}$.

To prove Proposition 5.3 we will require the following preliminary result.

LEMMA 5.3. *Under conditions* (RLS1)-(RLS4) *there exists a fixed constant* $B = B(E[|w_1|^2]) < \infty$ *such that*

$$E_{\Phi_0}[|x_{T_0+1}|^2] \leqq \lambda_0 |x_0|^2 + B$$

for every initial condition $\Phi_0 = \binom{x_0}{\xi_0} \in \mathbf{X}$.

Proof. We have for all $k \in \mathbb{Z}_+$ by (RLS2),

$$E_{\Phi_0}[|x_{k+1}|^2] = x_0^{\mathsf{T}} E_{\xi_0}[F(\xi_0)^{\mathsf{T}} \cdots F(\xi_k)^{\mathsf{T}} F(\xi_k) \cdots F(\xi_0)] x_0 + E_{\Phi_0}\left[w_{k+1}^{\mathsf{T}} w_{k+1} + \sum_{i=1}^{k} w_i^{\mathsf{T}} [F(\xi_i)^{\mathsf{T}} \cdots F(\xi_k)^{\mathsf{T}} F(\xi_k) \cdots F(\xi_i)] w_i \right].$$

Hence the conclusions of the lemma are satisfied with

$$B \triangleq E[|w_1|^2](T_0 \| F \|^{T_0}),$$

where $\|F\|$ denotes the supremum of the operator norm of F over the compact manifold M. □

We may now prove Proposition 5.3. We assume that an i.i.d. $N(0, I)$ stochastic process $\mathbf{d}$ on $\mathbb{R}^n$ exists such that $\mathbf{d}$, $\mathbf{w}$, and ξ are mutually independent. We then define the perturbed process $\mathbf{x}^\varepsilon$ for $\varepsilon \in [0, 1]$ by $x_0^\varepsilon = x_0$, and

$$x_{k+1}^\varepsilon = F(\xi_k) x_k^\varepsilon + w_{k+1} + \varepsilon d_{k+1}, \qquad k \in \mathbb{Z}_+. \tag{51}$$

To prove Proposition 5.3 we will show that for all $\varepsilon > 0$, the perturbed process is aperiodic and positive Harris recurrent, and that for ε sufficiently small, the process $\mathbf{x}^\varepsilon$ approximates the process $\mathbf{x}$, uniformly in $k \in \mathbb{Z}_+$.

Proof of Proposition 5.3. When $\varepsilon > 0$, the conditional probability on $\mathcal{B}(\mathbf{X})$ defined by

$$P_{(x_0, \xi_0)}\{x_{k+1}^\varepsilon \in A \,|\, \xi_1, \cdots, \xi_k\}$$

possesses a density that is continuous and everywhere positive on $\mathbb{R}^{2n} \times M^{k+1}$. If A is a petite set for ξ, then it follows that $A \times K$ is petite for any compact set K of positive Lebesgue measure. This together with an argument similar to the proof of Proposition 2.2 shows that $\mathbf{\Phi}^\varepsilon$ is Harris recurrent. Since Lemma 5.3 implies that condition (S2) holds, an invariant probability exists for all $\varepsilon > 0$ and we conclude that $\mathbf{\Phi}^\varepsilon$ is positive Harris recurrent for all $\varepsilon > 0$. The Markov chain $\mathbf{\Phi}^\varepsilon$ is aperiodic for $\varepsilon > 0$ by (RLS3), and since the distribution of x_k^ε possesses the same null sets as Lebesgue measure on $\mathbb{R}^n$ for all $k \geqq 1$.

From (51) and Lemma 5.3 there exists a constant $B < \infty$ such that

$$E_{\Phi_0}[|x_{T_0+1}^\varepsilon - x_{T_0+1}|^2] \leqq \frac{\varepsilon^2 B}{1 - \lambda_0}$$

for every initial condition $\Phi_0 \in \mathbf{X}$, and it follows that

$$\limsup_{\varepsilon \to 0} \sup_{k \geqq 0} E_{\Phi_0}[|x_k^\varepsilon - x_k|^2] = 0 \tag{52}$$

for all initial conditions.

Fix $f \in \mathbf{C}$ uniformly continuous. By (52) and Chebyshev's inequality

$$\limsup_{\varepsilon \to 0} \sup_{k \geqq 0} E_{\Phi_0}[|f(x_k^\varepsilon) - f(x_k)|] = 0, \tag{53}$$

and since $\mathbf{\Phi}^\varepsilon$ is aperiodic and positive Harris recurrent for each $\varepsilon > 0$, there exists an invariant probability π_ε on $\mathcal{B}(\mathbf{X})$ such that

$$\lim_{k \to \infty} E_{\Phi_0}[f(\Phi_k^\varepsilon)] = \int f \, d\pi_\varepsilon. \tag{54}$$

Using Lemma 5.3 it may be shown that the invariant probabilities $\{\pi_\varepsilon : 0 < \varepsilon \leqq 1\}$ possess uniformly bounded second moments, and hence are tight. We may therefore find a subsequence $\varepsilon_i \to 0$ as $i \to \infty$, and a probability π on $\mathcal{B}(\mathbf{X})$ such that

$$\pi_{\varepsilon_i} \xrightarrow{\text{weakly}} \pi \quad \text{as } i \to \infty. \tag{55}$$

Combining (53)-(55) we have for all $i \in \mathbb{Z}_+$,

$$\begin{aligned}
&\limsup_{k\to\infty} \left| E_{\Phi_0}[f(\Phi_k)] - \int f\,d\pi \right| \\
&\quad \leqq \limsup_{k\to\infty} |E_{\Phi_0}[f(\Phi_k)] - E_{\Phi_0}[f(\Phi_k^{\varepsilon_i})]| \\
&\qquad + \limsup_{k\to\infty} \left| E_{\Phi_0}[f(\Phi_k^{\varepsilon_i})] - \int f\,d\pi_{\varepsilon_i} \right| + \left| \int f\,d\pi_{\varepsilon_i} - \int f\,d\pi \right| \\
&\quad \leqq \sup_{j\in\mathbb{Z}_+} |E_{\Phi_0}[f(\Phi_j)] - E_{\Phi_0}[f(\Phi_j^{\varepsilon_i})]| + \left| \int f\,d\pi_{\varepsilon_i} - \int f\,d\pi \right|.
\end{aligned}$$

Letting $i \to \infty$ shows that

$$\lim_{k\to\infty} E_{\Phi_0}[f(\Phi_k)] = \int f\,d\pi.$$

This implies that π is an invariant probability for the unperturbed process, and also that π is the unique weak limit point of the probabilities $\{\pi_\varepsilon : \varepsilon > 0\}$. Hence $\pi_\varepsilon \xrightarrow{\text{weakly}} \pi$ as $\varepsilon \to 0$.

It follows that π is the unique invariant probability for $\mathbf{\Phi}$, and that for all $x \in \mathbf{X}$ the limit (48) holds. By a simple calculation

$$\sup_{k\geqq 0} E_{\Phi_0}[|x_k|^{2+\delta}] < \infty$$

for each deterministic initial condition, and this together with (48) implies that

$$\lim_{k\to\infty} E_x[x_k^2] = \int x^2\,d\pi < \infty.$$

To complete the proof of Proposition 5.3 we show that conditions (A2) and (A3) of § 4 are satisfied. Since we have shown that (A1) holds, result (49) will follow from Proposition 4.2. Since the proof is fairly routine, we will be brief.

By Lemma 5.3 there exists a constant $B < \infty$ such that

$$P^{T_0}|x|^2(x) \leqq \lambda_0 |x|^2 + B.$$

It is easily shown that the equation above implies condition (A2), and by a martingale difference sequence argument it may be shown that

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} |x_k|^{2+\delta/2} < \infty \quad \text{a.s. } [P_{\Phi_0}]$$

for all initial conditions, and in particular, condition (A3) holds.

This shows that the result (49) holds, and result (50) follows from (48), (49), and the previous inequality that implies that the function x^2 is uniformly integrable with respect to the occupation probabilities with probability one. □

5.3. Stochastic adaptive control. Consider the single-input single-output random parameter system model

$$y_{k+1} = \theta_k y_k + u_k + v_{k+1}, \qquad k \in \mathbb{Z}_+ \tag{56}$$

where the parameter process $\boldsymbol{\theta}$ is the output of the (AR1) model

$$\theta_{k+1} = \alpha\theta_k + e_{k+1}, \qquad k \in \mathbb{Z}_+ \tag{57}$$

and $\alpha \in (-1, 1)$.

The parameter process $\boldsymbol{\theta}$ is assumed to be unknown, but is estimated by the gradient algorithm

$$\hat{\theta}_{k+1} = \alpha\hat{\theta}_k + \alpha \frac{y_k(y_{k+1} - \hat{\theta}_k y_k - u_k)}{1 + y_k^2}. \tag{58}$$

This is a simplified version of the example analyzed in Meyn and Caines (1987) where $\hat{\boldsymbol{\theta}}$ is a version of the conditional expectation $\hat{\theta}_k = E[\theta_k \mid \mathcal{Y}_k]$, with $\mathcal{Y}_k \triangleq \sigma\{y_0, \cdots, y_k\}$. The estimator (58) was obtained by setting the estimation error covariance P_k in the algorithm of Meyn and Caines (1987) to a constant. Hence in the present example the parameter estimates $\{\hat{\theta}_k\}$ have no simple interpretation.

Assume now that our goal is to choose a control $u_k \in \mathcal{Y}_k$ that makes the mean square output error $E[y_k^2]$ as small as possible. When we apply the certainty equivalence control law

$$u_k = -\hat{\theta}_k y_k, \qquad k \in \mathbb{Z}_+ \tag{59}$$

and defining $\tilde{\theta}_k \triangleq \theta_k - \hat{\theta}_k$, the closed loop system equations become

$$\Phi_{k+1} \triangleq \begin{pmatrix} y_{k+1} \\ \tilde{\theta}_{k+1} \end{pmatrix} = \begin{pmatrix} \tilde{\theta}_k y_k + v_{k+1} \\ \alpha \dfrac{\tilde{\theta}_k - y_k v_{k+1}}{1 + y_k^2} + e_{k+1} \end{pmatrix}, \qquad k \in \mathbb{Z}_+. \tag{60}$$

It is evident that, under the appropriate conditions on the process $\mathbf{w} \triangleq \binom{\mathbf{v}}{\mathbf{e}}$, the state process $\boldsymbol{\Phi}$ is a Feller Markov chain of the form (1) with state space $\mathbf{X} \triangleq \mathbb{R}^2$.

We henceforth assume that $\mathbf{w}$ satisfies (W1) and (W2) that $\mathbf{v}$ and $\mathbf{e}$ are independent, and that the following additional assumptions hold for some $\delta > 0$:

$$E[w_1] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad E[|w_1|^{4+\delta}] < \infty, \quad E[|e_1|^{2+\delta}] < 1. \tag{61}$$

These conditions imply that $\sigma_e^2 \triangleq E[|e_1|^2] < 1$, $\sigma_v^2 \triangleq E[|v_1|^2] < \infty$, and $\gamma_v^4 \triangleq E[|v_1|^2] < \infty$.

The state process $\boldsymbol{\Phi}$ is weakly stochastically controllable, and satisfies condition (GA) with $d^* = 0$. By Proposition 5.1 we conclude that the irreducibility hypothesis holds for an open petite set $A \subset \mathbf{X}$ and that $\boldsymbol{\Phi}$ is aperiodic. Our next task is to find a moment satisfying (12) so that we may apply Proposition 3.1.

Let $y : \mathbf{X} \to \mathbb{R}$ and $\tilde{\theta} : \mathbf{X} \to \mathbb{R}$ denote the coordinate variables on $\mathbf{X}$ so that

$$y_k = y(\Phi_k), \quad \tilde{\theta}_k = \tilde{\theta}(\Phi_k) \quad \text{for all } k \in \mathbb{Z}_+,$$

and define the test function V on $\mathbf{X}$ by

$$V(y, \tilde{\theta}) = \tilde{\theta}^4 + \varepsilon_0 \tilde{\theta}^2 y^2 + \varepsilon_0^2 y^2 \tag{62}$$

where $\varepsilon_0 > 0$ is a small constant which will be specified below.

Letting P denote the Markov transition function for Φ we have by (60),

$$Py^2 = \tilde{\theta}^2 y^2 + \sigma_v^2. \tag{63}$$

This is far from (12), but applying the operator P to the function $\tilde{\theta}^2 y^2$ gives

$$\begin{aligned} P\tilde{\theta}^2 y^2 &= E\left[\left(\frac{\alpha\tilde{\theta} - \alpha y v_1}{1+y^2} + e_1\right)^2 (\tilde{\theta} y + v_1)^2\right] \\ &= \sigma_e^2 \tilde{\theta}^2 y^2 + \sigma_e^2 \sigma_v^2 + \left(\frac{\alpha}{1+y^2}\right)^2 E[(\tilde{\theta} - y v_1)^2 (\tilde{\theta} y + v_1)^2] \\ &= \sigma_e^2 \tilde{\theta}^2 y^2 + \sigma_e^2 \sigma_v^2 + \left(\frac{\alpha}{1+y^2}\right)^2 [\tilde{\theta}^4 y^2 + \tilde{\theta}^2 \sigma_v^2 - 4\tilde{\theta}^2 y^2 \sigma_v^2 + \tilde{\theta}^2 y^4 \sigma_v^2 + y^2 \gamma_v^4], \end{aligned}$$

and hence we may find a constant $B_1 < \infty$ such that

$$P\tilde{\theta}^2 y^2 \leqq \sigma_e^2 \tilde{\theta}^2 y^2 + B_1(\tilde{\theta}^4 + \tilde{\theta}^2 + 1). \tag{64}$$

From (60) it is easy to show that for some constant $B_2 > 0$

$$P\tilde{\theta}^4 \leqq \alpha^4 \tilde{\theta}^4 + B_2(\tilde{\theta}^2 + 1). \tag{65}$$

Combining equations (63)-(65) we may find a constant $K_3 < \infty$, such that for all $0 < \varepsilon < 1$,

$$\begin{aligned} P(\tilde{\theta}^4 + \varepsilon\tilde{\theta}^2 y^2 + \varepsilon^2 y^2) &\leqq (\alpha^4 + \varepsilon K_3)\tilde{\theta}^4 + (\sigma_e^2 + \varepsilon)\varepsilon\tilde{\theta}^2 y^2 + K_3(\tilde{\theta}^2 + 1) \\ &\leqq (\alpha^4 + 2\varepsilon K_3)\tilde{\theta}^4 + (\sigma_e^2 + \varepsilon)\varepsilon\tilde{\theta}^2 y^2 + 2K_3/\varepsilon \end{aligned} \tag{66}$$

where the second inequality follows from the estimate $\tilde{\theta}^2 \leqq \varepsilon\tilde{\theta}^4 + 1/\varepsilon$. Fix $1 > \lambda > \max(\sigma_e^2, \alpha^4)$. Then by (66) we may find $\varepsilon_0 > 0$ sufficiently small, and a constant $K > 0$ sufficiently large such that (12) holds with V defined in (62).

A modification of (63) and (64) may be used together with (61) to show that

$$\sup_{k \geqq 0} E_x[|y_k|^{2+\delta}] < \infty,$$

and applying Proposition 3.1 and Proposition 2.1 we obtain Proposition 5.4.

PROPOSITION 5.4. *The Markov chain Φ defined in* (60) *is geometrically ergodic, and for all initial conditions $x \in \mathbf{X}$,*

$$\lim_{k\to\infty} E_x[y_k^2] = \int y^2 \, d\pi$$

at a geometric rate, and

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} y_k^2 = \int y^2 \, d\pi < \infty \quad a.s.\,[P_x].$$

6. Conclusions. The principal tool used in both stochastic and deterministic stability theory is some form of Lyapunov function. In this paper we have taken a specific Lyapunov function that at first sight appears to provide at most a very crude form of stability, but in fact has broad implications to ergodic theory and implies stronger stability results that might be expected.

We have shown that under general conditions its existence implies that the law of large numbers holds for a large class of functions of the Markov chain Φ and all initial conditions. Furthermore, if a stochastic controllability condition is satisfied, then the underlying distributions governing the system converge to an invariant

probability at a geometric rate. This is known as geometric ergodicity in the Markov chain literature, but may also be seen as a form of exponential asymptotic stability of the Markov transition operator P acting on $\mathcal{M}$ = the set of all probabilities on $\mathcal{B}(\mathbf{X})$.

There are at least two implications of these results that require further study. First of all, it appears that the results presented here will find application to the ODE (ordinary differential equation) method (see, for example, Kushner (1983)). One of the key hypotheses of Kushner (1983) is the existence of limits of the form

$$\lim_{N\to\infty} \sum_{k=n}^{N} a_k (P^{k-n} G(\Phi_n) - \bar{G})$$

where a_k is a square summable scalar sequence. We feel that geometric ergodicity is an obvious route to proving that this limit exists and computing bounds on the limit. If this is the case, then test functions satisfying (12) should be a powerful tool when combined with the ODE method. This idea has already been pursued in Arapostathis and Marcus (1988).

Another possible application is to the structural robustness of stochastic systems. It is well known that the solutions of the ordinary differential equation

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

will be uniformly bounded if the error term ε is sufficiently small in some sense, and the "ideal system" $\dot{x} = f(x)$ is sufficiently stable.

By considering the dynamical system on $\mathcal{M}$ generated by a Markov transition operator P, and using the fact that geometric ergodicity is simply a form of exponential asymptotic stability for P, it may be possible to extend this result to the stochastic case.

Acknowledgments. I thank Professor Peter Caines of McGill University for a number of stimulating discussions on stability theory at the time that I was writing § 3, and for suggesting the example on nonlinear control.

REFERENCES

A. ARAPOSTATHIS AND MARCUS (1988), *On the adaptive control of a partially observable Markov decision process*, in Proc. 27th IEEE Conference on Decision and Control, Austin, Texas.

K. B. ATHREYA AND P. NEY (1980), *Some aspects of ergodic theory and laws of large numbers for Harris recurrent Markov chains*, Colloquia Mathematica Societatis János Bolyai, 32, Nonparametric Statistical Inference, Budapest, Hungary, pp. 41-56.

V. E. BENEŠ (1968), *Finite regular invariant measures for Feller processes*, J. Appl. Probab., 5, pp. 203-209.

P. BILLINGSLEY (1968), *Convergence of Probability Measures*, John Wiley, New York.

L. BREIMAN (1960), *The strong law of large numbers for a class of Markov chains*, Ann. Math. Statist., 31, pp. 801-803.

K. S. CHAN (1986), *Topics in nonlinear time series analysis*, Ph.D. thesis, Department of Mathematics, Princeton University, Princeton, NJ.

R. COGBURN (1975), *A uniform theory for sums of Markov chain transition probabilities*, Ann. Probab., 3, pp. 191-214.

J. L. DOOB (1953), *Stochastic Processes*, John Wiley, New York.

S. R. FOGUEL (1969), *Positive operators on $C(X)$*, Proc. Amer. Math. Soc., 22, pp. 295-297.

D. F. FEIGIN AND R. L. TWEEDIE (1985), *Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments*, J. Time Ser. Anal., 6, pp. 1-14.

W. FELLER (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, John Wiley, New York.

F. G. FOSTER (1953), *On the stochastic matrices associated with certain queueing processes*, Ann. of Math. Statist., 24, pp. 355-360.

G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES (1981), *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19, pp. 829-853; Corrigendum, 20 (1982), p. 893.

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., Vol. 3, pp. 1-14.

R. Z. HAS'MINSKIĬ (1980), *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff Alphen an den Rijn, the Netherlands, Rockville, Maryland.

B. JAKUBCZYK AND E. D. SONTAG (1990), *Controllability of nonlinear discrete time systems: a Lie-algebraic approach*, SIAM J. Control Optim., 28, to appear.

R. E. KALMAN AND J. E. BERTRAM (1960), *Control system analysis and design by the second method of Lyapunov*, Trans. ASME Ser. D: J. Basic Engrg., 82, pp. 371–400.

H. J. KUSHNER (1983), *An averaging method for stochastic approximations with discontinuous dynamics, constraints, and state dependent noise*, in Recent Advances in Statistics, M. H. Rizvi, J. Rustagi, and D. Siegmund, eds., Academic Press, New York, pp. 211–235.

———, (1972), *Stochastic Stability*, Lecture Notes in Mathematics 294, Springer-Verlag, New York.

———, (1967), *Stochastic Stability and Control*, Academic Press, New York.

S. P. MEYN (1987), *Asymptotic behavior of stochastic systems possessing Markovian realizations*, Ph.D. thesis, Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada.

S. P. MEYN AND P. E. CAINES (1987), *A new approach to stochastic adaptive control*, IEEE Trans. Automat. Control, 32, pp. 220–226.

———, (1988), *Asymptotic behavior of stochastic systems possessing Markovian realizations*, SIAM J. Control Optim., submitted.

E. NUMMELIN (1984), *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, Cambridge.

E. NUMMELIN AND P. TUOMINEN (1982), *Geometric ergodicity of Harris recurrent Markov chains with applications to renewal theory*, Stochastic Process. Appl., 12, pp. 187–202.

S. OREY (1971), *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold Mathematical Studies 34, Van Nostrand, London.

K. R. PARTHASARATHY (1967), *Probability Measures on Metric Spaces*, Academic Press, New York.

V. M. POPOV (1973), *Hyperstability of Control Systems*, Springer-Verlag, Berlin, and Editura Academici Bucuresti.

D. REVUZ (1975), *Markov Chains*, North-Holland, Amsterdam.

M. ROSENBLATT (1971), *Markov Processes: Structure and Asymptotic Behaviour*, Springer-Verlag, Berlin, Heidelberg, New York.

M. G. SAFONOV (1980), *Stability and Robustness of Multivariable Feedback Systems*, MIT Press, Cambridge, MA.

V. SOLO (1978), *Time series recursions and stochastic approximation*, Ph.D. dissertation, The Australian National University, Canberra, Australia, September 1978.

P. TUOMINEN AND R. L. TWEEDIE (1979), *Markov chains with continuous components*, Proc. London Math. Soc. (3), 38, pp. 89–114.

R. L. TWEEDIE (1983), *Criteria for rates of convergence of Markov chains, with application to queueing and storage theory*, in Probability, Statistics and Analysis, London Math. Society Lecture Note Series 79, J. F. C. Kingman and G. E. H. Reuter, eds., Cambridge University Press, Cambridge.

———, (1976), *Criteria for classifying general Markov chains*, Adv. in Appl. Probab., 8, pp. 737–771.

———, (1975), *Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space*, Stochastic Process. Appl., 3, pp. 385–403.

G. ZAMES (1966a), *On the input-output stability of time varying nonlinear feedback systems*, Part I: IEEE Trans. Automat. Control, 11, pp. 228–238.

———, (1966b), *On the input-output stability of time varying nonlinear feedback systems, Part* II, IEEE Trans. Automat. Control, 11, pp. 465–476.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1440–1461, November 1989

012

SIMULATED ANNEALING TYPE MARKOV CHAINS AND THEIR ORDER BALANCE EQUATIONS*

DANIEL P. CONNORS† AND P. R. KUMAR‡

Abstract. Generalized simulated-annealing type Markov chains where the transition probabilities are proportional to powers of a vanishing small parameter are considered. An "order of recurrence," which quantifies the asymptotic behavior of the state occupation probability, is associated with each state. These orders of recurrence satisfy a fundamental balance equation across every edge-cut in the graph of the Markov chain. Moreover, the Markov chain converges in a Cesaro-sense to the set of states having the largest recurrence orders. These results convert the analytic problem of determining the asymptotic properties of the time-inhomogeneous stochastic process into a purely algebraic problem of solving the balance equations to determine the recurrence orders.

Graph theoretic algorithms are provided to determine the solutions of the balance equations. By applying these results to the problem of optimization by simulated annealing, it is shown that the sum of the recurrence order and the cost is a constant for all states in a certain connected set, whenever a "weak-reversibility" condition is satisfied. This allows the necessary and sufficient condition for the optimization algorithm to hit the global minimum with probability one to be obtained.

Key words. simulated annealing, optimization, Markov chains

AMS(MOS) subject classifications. 60J10, 90C27

1. Introduction. We consider finite state Markov chains $\{x(t)\}$ with transition probabilities of the type

$$p_{ij}(t) = c_{ij}\varepsilon(t)^{V_{ij}},$$

where $\varepsilon(t)$ is a small parameter converging to zero. In a previous paper [7] we have shown that if we define "orders of recurrence" by (more precise definitions are given in § 2)

$$\beta_i := \sup\left\{c \geqq 0: \sum_{t=0}^{\infty} \varepsilon(t)^c \pi_i(t) = +\infty\right\},$$

then

(i) These recurrence orders satisfy a balance equation, $\max_{i\in A, j\in A^c}(\beta_i - V_{ij}) = \max_{i\in A, j\in A^c}(\beta_j - V_{ji})$, for every subset A; and

(ii) The Markov process converges to the set of states with the largest orders of recurrence.

This provides a novel approach to analyzing the asymptotic behavior of such time-inhomogeneous Markov processes. Specifically, we use (i) to solve the balance equations, and then (ii) provides the limiting behavior. Moreover, the orders of recurrence also provide information about the rates of convergence of the state occupation probabilities. This approach via recurrence orders therefore converts the analytic problem of determining the asymptotic behavior of the time-inhomogeneous process into a purely algebraic problem of solving the balance equations.

* Received by the editors July 11, 1988; accepted for publication (in revised form) December 30, 1988.
This research has been supported in part by Air Force Office of Scientific Research contract AFOSR-88-0181, U.S. Army Research Office contract DAAL-03-88-K0046, and Joint Services Electronics Program contract N00014-84-C-0149.

† IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598.

‡ Department of Electrical and Computer Engineering, and the Coordinated Science Laboratory, University of Illinois, 1101 W. Springfield Avenue, Urbana, Illinois 61801.

A significant motivation for studying such Markov chains lies in the fact that in the method of optimization by simulated annealing, if $\{W_i\}$ is the cost function whose minimum is sought, then we obtain a Markov chain with

$$p_{ij}(t) = c_{ij}\varepsilon(t)^{\max(0, W_j - W_i)}.$$

Thus simulated annealing is a special case where the powers V_{ij} satisfy

$$V_{ij} := \max(0,\ W_j - W_i),$$

for some $\{W_i\}$.

To pursue the above approach to analyzing such time-inhomogeneous Markov chains, it is necessary to be able to solve the balance equations. However, there can be nonunique solutions to the balance equations. We present graph-theoretic circulation based algorithms to obtain *a* solution, as well as *all* solutions, to the balance equations. We show by an example the interesting phenomenon that such nonuniqueness can arise when the asymptotic properties of the Markov process, and the recurrence orders, depend not just on the *exponents* V_{ij}, but also on the *proportionality constants* c_{ij}.

By applying these results to the Markov chain arising from the method of optimization by simulated annealing when the "weak reversibility" condition of Hajek [1] holds, we show that the sum of the recurrence order and the cost is a constant on sets connected by recurrent arcs. This allows us to obtain the necessary and sufficient condition for the optimization algorithm to hit the global minimum with probability one. Our necessity result is a stronger sample path result than is found in [1] or [2].

Background. Tsitsiklis [2] has also investigated Markov chains with transition probabilities proportional to powers of a small time-varying parameter. His analysis was based on observing that due to the slow variation of $\{\varepsilon(t)\}$, we can employ bounds on the state occupation probabilities for *stationary* Markov chains, where $\varepsilon(t)$ is held constant, to obtain bounds for the time-inhomogeneous case. His approach is quite different from ours.

Based on an analogy to the physical process of annealing, the sequence $\varepsilon(t)$ is called the "cooling schedule," and just as in the physical analogy it plays a key role in determining asymptotic behavior. It has been shown by Geman and Geman [3], Mitra, Romeo, and Sangiovanni-Vincentelli [4], and Gidas [5], that simulated annealing converges in probability to a minimum of the optimization problem provided $\sum_{t=0}^{\infty} \varepsilon(t)^p = +\infty$ for large enough p. Hajek [1] has determined the necessary and sufficient conditions on the value of p for the algorithm to converge *in probability* to the minimum when a "weak reversibility" assumption is satisfied.

2. Orders of recurrence and balance equations. Consider a Markov chain over a finite state space X whose transition probabilities are proportional to *powers* of a vanishing time varying parameter $\varepsilon(t)$; that is, the transition probabilities $p_{ij}(t) := \Pr(x(t+1) = j \mid x(t) = i)$ are given by

$$(1) \qquad p_{ij}(t) = c_{ij}\varepsilon(t)^{V_{ij}} \quad \text{for all } i, j \in X,\ i \neq j, \text{ and } t \in \mathscr{Z}^+, \text{ and } p_{ii}(t) = 1 - \sum_{j \neq i} p_{ij}(t)$$

where

$$(2) \qquad 0 \leqq V_{ij} \leqq +\infty \quad \text{for all } i, j \in X,\ i \neq j,$$

$$(3) \qquad c_{ij} \geqq 0 \quad \text{for all } i, j \in X,\ i \neq j, \text{ and } \sum_j c_{ij} = 1 \text{ for all } i.$$

Regarding the small parameter $\{\varepsilon(t)\}$, we will assume that,

$$0<\varepsilon(t)<1 \quad \text{for all } t\in\mathscr{Z}^+, \tag{4}$$

$$\exists M<\infty \text{ such that } \varepsilon(t)\leqq M\varepsilon(s) \text{ whenever } t\geqq s, \text{ and} \tag{5}$$

$$\sum_{t=1}^{\infty}\varepsilon(t)^p<\infty \quad \text{for some } p\in[1,+\infty). \tag{6}$$

In what follows we will assume that in (1)-(3) we have

$$c_{ij}=0\Leftrightarrow V_{ij}=+\infty,$$

which is clearly without any loss of generality. We shall denote by N_i the set of all states j with $c_{ij}>0$. Finally, we will assume that the Markov chain is "connected;" i.e., for every $i,j\in X$, there exists a path $i=i_0,\cdots,i_p=j$, with $i_l\in N_{i_{l-1}}$ for $1\leqq l\leqq p$.

Let $\pi_i(t):=\Pr(x(t)=i)$ be the probability distribution of $x(t)$, and let $\pi_{ij}(t):=\Pr(x(t)=i, x(t+1)=j)$ be the probability of a *transition* from state i to j at time t.

The following example motivates the notion of "orders of recurrence" introduced in [7].

Example 1. Suppose, for a certain Markov chain (with more than two states!), we have

$$\pi_1(t)=1/t^{1/3},\quad \pi_2(t)=1/t^{2/3},\quad \varepsilon(t)=1/t^{1/3}.$$

Then note that $\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_1(t)$ is *finite* if $c>\beta_1:=2$ and $+\infty$ if $c\leqq\beta_1$. Similarly, $\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_2(t)$ is *finite* if $c>\beta_2:=1$ and $+\infty$ if $c\leqq\beta_2$. Now $\pi_1(t)$ converges to zero more slowly than $\pi_2(t)$ and it is easy to see that this information is also captured by the demarcation points β_1 and β_2, which thus provide a measure by which to rank the rates at which $\pi_1(t)$ and $\pi_2(t)$ converge to zero.

Motivated by this we define the *recurrence orders* for the states and transitions of the Markov process, as follows.

DEFINITION 1. The order of recurrence of a state $i\in X$, denoted β_i, is

$$\beta_i:=\begin{cases}-\infty & \text{if } \sum_{t=0}^{\infty}\pi_i(t)<+\infty,\\ p^- & \text{if } p=\sup\left\{c\geqq 0:\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_i(t)=+\infty\right\} \text{ and } \sum_{t=0}^{\infty}\varepsilon(t)^p\pi_i(t)<+\infty,\\ p & \text{if } p=\max\left\{c\geqq 0:\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_i(t)=+\infty\right\}.\end{cases}$$

We say a state i is *transient* if $\beta_i=-\infty$; otherwise we say the state is *recurrent.*

In a similar manner we define the *order of recurrence of the transition from i to j.*

DEFINITION 2. The order of recurrence of the transition from state i to j, denoted β_{ij}, is

$$\beta_{ij}:=\begin{cases}-\infty & \text{if } \sum_{t=0}^{\infty}\pi_{ij}(t)<+\infty,\\ p^- & \text{if } p=\sup\left\{c\geqq 0:\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_{ij}(t)=+\infty\right\} \text{ and } \sum_{t=0}^{\infty}\varepsilon(t)^p\pi_{ij}(t)<+\infty,\\ p & \text{if } p=\max\left\{c\geqq 0:\sum_{t=0}^{\infty}\varepsilon(t)^c\pi_{ij}(t)=+\infty\right\}.\end{cases}$$

Again, we say the transition from i to j is *transient* if $\beta_{ij}=-\infty$; otherwise we say the transition is *recurrent.*

It is also convenient to define ρ, the *order of cooling* of $\{\varepsilon(t)\}$, as follows.

DEFINITION 3. The order of the cooling schedule $\{\varepsilon(t)\}$, denoted ρ, is defined as

$$\rho := \begin{cases} -\infty & \text{if } \sum_{t=0}^{\infty} \varepsilon(t) < +\infty, \\ p^- & \text{if } p = \sup\left\{c \geqq 0: \sum_{t=0}^{\infty} \varepsilon(t)^c = +\infty\right\} \text{ and } \sum_{t=0}^{\infty} \varepsilon(t)^p < +\infty, \\ p & \text{if } p = \max\left\{c \geqq 0: \sum_{t=0}^{\infty} \varepsilon(t)^c = +\infty\right\}. \end{cases}$$

The relationship between β_i, β_{ij}, and ρ is given in the following lemma. It will be convenient in the sequel to define the operation "$\ominus$" as follows:

$$a \ominus b := \begin{cases} -\infty & \text{if } a < b, \\ a - b, & \text{if } a \geqq b. \end{cases}$$

LEMMA 1. *β_{ij} and β_i are related by*

$$\beta_{ij} = \beta_i \ominus V_{ij} \quad \textit{for all } i, j \in X, \tag{7}$$

while ρ and β_i are related by

$$\max_{i \in X} \beta_i = \rho. \tag{8}$$

Proof. If $j \notin N_i$, then it immediately follows that $\beta_{ij} = -\infty$. If $j \in N_i$, then application of the Chapman-Kolmogorov equation

$$\begin{aligned} \pi_{ij}(t) &= \pi_i(t) p_{ij}(t) \\ &= c_{ij} \varepsilon(t)^{V_{ij}} \pi_i(t), \end{aligned}$$

gives the first assertion. Similarly, since

$$\sum_{t=0}^{\infty} \varepsilon(t)^p = \sum_{i \in X} \sum_{t=0}^{\infty} \varepsilon(t)^p \pi_i(t),$$

the second assertion also follows. □

Knowledge of the β_i's provides useful information about the asymptotic properties of $\{x(t)\}$. The following theorem shows that the time-inhomogeneous Markov chain converges in a *Cesaro sense* to the set of states having the largest orders of recurrence.

THEOREM 1. *Let $\mathcal{M}$ be the set of states with the largest orders of recurrence*:

$$\mathcal{M} := \{i \in X: \beta_i = \rho\}.$$

Then

$$\limsup_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \Pr(x(t) \in \mathcal{M}) = 1. \tag{9}$$

Proof. Let us first consider the set $\bar{\mathcal{M}}$ defined by

$$\bar{\mathcal{M}} := \begin{cases} \mathcal{M} & \text{if } \rho = 0, -\infty \text{ or } p^- \text{ for some } p \in \mathcal{R},\ p > 0, \\ \mathcal{M} \cup \{i \in X: \beta_i = p^-\} & \text{if } \rho = p \text{ for some } p \in \mathcal{R},\ p > 0. \end{cases}$$

Note that if $\rho = p$, then $\bar{\mathcal{M}}$ may be slightly larger than $\mathcal{M}$ since it includes states, if any, whose recurrence orders are p^-; otherwise it is the same as $\mathcal{M}$. We will first show that

$$\limsup_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \Pr(x(t) \in \bar{\mathcal{M}}) = 1. \tag{10}$$

Consider first the case $\rho > 0$. Clearly, $\rho = p$ or p^- for some $p \in \mathcal{R}$, where $p > 0$. Let

$$Q = \{q \in \mathcal{R}: \text{ for some } i \in \bar{\mathcal{M}}^c,\ \beta_i = q \text{ or } q^-\}.$$

Let $\theta = \inf_{q \in Q} (p - q)$, where $\inf \varnothing = +\infty$. Let

$$\gamma = \begin{cases} \theta & \text{if } \theta < +\infty, \\ p & \text{if } \theta = +\infty. \end{cases}$$

Consider the states in $\bar{\mathcal{M}}^c$ and observe that for sufficiently small $\delta > 0$,

$$\sum_{t=0}^{\infty} \Pr(x(t) \in \bar{\mathcal{M}}^c) \varepsilon(t)^{p-\gamma+\delta} < +\infty,$$

since the state space is finite. An application of Kronecker's Lemma (see Chung [6]) gives

$$\lim_{N \to \infty} \varepsilon(N)^{p-\gamma+\delta} \sum_{t=1}^{N} \Pr(x(t) \in \bar{\mathcal{M}}^c) = 0;$$

that is,

$$\lim_{N \to \infty} (N\varepsilon(N)^{p-\gamma+\delta}) \frac{1}{N} \sum_{t=1}^{N} \Pr(x(t) \in \bar{\mathcal{M}}^c) = 0. \tag{11}$$

Now we claim that

$$\limsup_{N \to \infty} N\varepsilon(N)^{p-\gamma+\delta} > 0. \tag{12}$$

Suppose not. Then,

$$\lim_{N \to \infty} N\varepsilon(N)^{p-\gamma+\delta} = 0,$$

and so

$$\lim_{N \to \infty} \frac{1/N}{\varepsilon(N)^{p-\gamma+\delta}} = +\infty.$$

In particular, we have

$$\lim_{N \to \infty} \left(\frac{1/N}{\varepsilon(N)^{p-\gamma+\delta}} \right)^{(p-\delta)/(p-\gamma+\delta)} = +\infty,$$

implying that

$$\lim_{N \to \infty} \frac{(1/N)^{(p-\delta)/(p-\gamma+\delta)}}{\varepsilon(N)^{p-\delta}} = +\infty.$$

However, since $\sum_{t=0}^{\infty} \varepsilon(t)^{p-\delta} = +\infty$, this would imply that

$$\sum_{N=1}^{\infty} \left(\frac{1}{N} \right)^{(p-\delta)/(p-\gamma+\delta)} = +\infty \quad \text{for all small } \delta > 0,$$

which is false. Hence, (12) holds and from (11) we deduce that

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \Pr(x(t) \in \bar{\mathcal{M}}^c) = 0. \tag{13}$$

But since

$$\frac{1}{N}\sum_{t=1}^{N}\Pr\left(x(t)\in\bar{\mathcal{M}}\right)+\frac{1}{N}\sum_{t=1}^{N}\Pr\left(x(t)\in\bar{\mathcal{M}}^c\right)=1,$$

the result (10) follows.

Now turn to the case $\rho=0$. Then clearly, $\sum_{t=0}^{\infty}\Pr\left(x(t)\in\bar{\mathcal{M}}^c\right)<+\infty$, and so (13) is again true and the result (10) follows.

If $\rho=-\infty$, the result (10) is trivial.

To proceed from (10) to (9), it is clearly sufficient to show that in the case $\rho=p$ for some $p\in\mathcal{R}$, $p>0$,

$$\lim_{t\to\infty}\Pr\left(x(t)\in\{i\colon \beta_i=p^-\}\right)=0.$$

This involves some results on the structure of the recurrence orders and is demonstrated in Lemma 5. □

Thus, knowledge of the recurrence orders $\{\beta_i\}$ provides knowledge about the asymptotic properties of the time-inhomogeneous Markov chain. In fact, as the reader may see from Example 1, the recurrence orders also provide information about the rates of convergence.

Our goal therefore is to determine the recurrence orders, and critical to that will be the following result established in [7], which shows that there is a fundamental balance of recurrence orders across every edge-cut in the graph of the Markov chain.

THEOREM 2 (Order Balance).

$$\max_{i\in A,j\in A^c}\beta_{ij}=\max_{i\in A,j\in A^c}\beta_{ji}\quad \textit{for every } A\subseteq X. \tag{14}$$

Equivalently, using the "⊖" notation and (7),

$$\max_{i\in A,j\in A^c}\beta_i\ominus V_{ij}=\max_{i\in A,j\in A^c}\beta_j\ominus V_{ji}\quad \textit{for every } A\subseteq X. \tag{15}$$

Proof. We sketch the proof; see [7] for the precise proof. Choose $A\subseteq X$ and note that if $\{\tau(n)\}_{n\geqq 1}$ is the sequence of random times at which the process moves from A to A^c, while $\{\sigma(n)\}_{n\geqq 1}$ is the sequence of random times at which the process moves from A^c back to A, then we have

$$\tau(n)<\sigma(n)<\tau(n+1),$$

where we have assumed, without loss of generality, that $x(0)\in A$ to give $\tau(1)<\sigma(1)$. Using this it follows from (5) that

$$\begin{aligned}
\sum_{t=0}^{+\infty}\varepsilon(t)^c I(x(t)\in A^c, x(t+1)\in A) &= \sum_{n=1}^{+\infty}\varepsilon(\sigma(n))^c\\
&\leqq M^c\sum_{n=1}^{+\infty}\varepsilon(\tau(n))^c\\
&= M^c\sum_{t=0}^{+\infty}\varepsilon(t)^c I(x(t)\in A, x(t+1)\in A^c)\\
&= M^c\sum_{n=1}^{+\infty}\varepsilon(\tau(n+1))^c + M^c\varepsilon(\tau(1))^c\\
&\leqq M^{2c}\sum_{n=1}^{+\infty}\varepsilon(\sigma(n))^c + M^{2c}\varepsilon(0)^c\\
&= M^{2c}\sum_{t=0}^{+\infty}\varepsilon(t)^c I(x(t)\in A^c, x(t+1)\in A) + M^{2c}\varepsilon(0)^c.
\end{aligned}$$

By taking expected values and using the Monotone Convergence Theorem, it follows that

$$\sum_{t=0}^{\infty} \varepsilon(t)^c \sum_{i\in A, j\in A^c} \pi_{ij}(t) < +\infty \Leftrightarrow \sum_{t=0}^{\infty} \varepsilon(t)^c \sum_{i\in A^c, j\in A} \pi_{ij}(t) < +\infty.$$

Hence both sides above converge or diverge together. Now if c is so large that every term on the left-hand side with $i \in A$, $j \in A^c$ converges, then clearly c is also so large that every term on the right-hand side converges. Thus,

$$c > \max_{i\in A, j\in A^c} \beta_{ij} \Leftrightarrow c > \max_{i\in A, j\in A^c} \beta_{ji}.$$

Likewise if c is small enough so that some term on the left-hand side diverges, then c is also small enough so that some term on the right-hand side diverges, and so

$$c \leqq \max_{i\in A, j\in A^c} \beta_{ij} \Leftrightarrow c \leqq \max_{i\in A, j\in A^c} \beta_{ji}. \qquad \square$$

Note that through Theorems 1 and 2 we have converted the problem of determining the asymptotic properties of the time-inhomogeneous Markov chain into an *algebraic* problem of solving the balance equations (14). Note that (14) provides a maximum of $2^{|X|}$ equations, one for each edgecut.

3. The modified balance equations. Note that if $(\beta_1, \beta_2, \cdots, \beta_{|X|})$ satisfy (15), then $(\beta_1 - a, \beta_2 - a, \cdots, \beta_{|X|} - a)$ also satisfy (15) for every a, i.e., the solution set is *translation invariant.* Thus (8), which fixes the *maximum* of the β_i's, also needs to be taken into account.

However, (15), (8) together can *still* possess nonunique solutions for sufficiently small values of ρ. In this section, we will show how we can obtain *one* solution to (15), (8); in the next section we show how to obtain *all* solutions.

In cases where there is a unique solution to the order balance equations, the algorithm of this section gives an $O(|X|^3)$ algorithm for determining it, compared to the algorithm of § 4 for obtaining all solutions (in the nonunique case), which is exponential in $|X|$. Also, the results of this section are used in the analysis of the simulated annealing algorithm in § 5.

It is convenient to consider the following "modified" balance equations that, as we show in the sequel, always possess a unique solution. Given $\rho \geqq 0$ and $V_{ij} \geqq 0$ for $i, j = 1, \cdots, |X|$ with $i \neq j$, consider the problem of determining $\lambda := (\lambda_1, \cdots, \lambda_{|X|})$ such that

$$\max_{i\in A, j\in A^c} \lambda_i - V_{ij} = \max_{i\in A, j\in A^c} \lambda_j - V_{ji} \quad \text{for every } A \subseteq \{1, \cdots, |X|\}, \tag{16}$$

and

$$\max_i \lambda_i = \rho. \tag{17}$$

We call (16), (17) the "modified" balance equations. Observe that (16) differs from (15) in that the operation "$-$" is used in place of "$\ominus$." Also, the λ's can be negative in (16).

We have introduced the modified balance equations to avoid the difficulties in handling $-\infty$ that occur under the "$\ominus$" operation.

THEOREM 3 (Properties of Order Balance and Modified Balance Equations). (1) *If λ satisfies the modified balance equations for a given ρ and V, then β defined by*

$$\beta_i := \lambda_i \ominus 0 \tag{18}$$

satisfies the order balance equations (15), (8) *for the given ρ and V.*

(2) *For every given ρ and V, there exists a unique solution λ to the modified balance equations. Moreover, the solutions for different values of ρ are translates of each other.*

(3) *Whenever ρ is large enough, there exists a unique solution to the order balance equations* (15), (8). *These unique solutions are all translates of the solutions for the modified balance equations.*

Proof. Suppose that for a fixed ρ and V, there exist *two distinct* solutions β and $\hat{\beta}$ to the order balance equations. Define

$$A := \{k \in X: \hat{\beta}_k \leqq \beta_k\}.$$

Then we claim that

$$\max_{i \in A, j \in A^c} \beta_i \ominus V_{ij} = \max_{i \in A, j \in A^c} \beta_j \ominus V_{ji} = -\infty$$

and

$$\max_{i \in A, j \in A^c} \hat{\beta}_i \ominus V_{ij} = \max_{i \in A, j \in A^c} \hat{\beta}_j \ominus V_{ji} = -\infty.$$

We need only consider the case where $A \neq \varnothing$ and $A \neq X$ (otherwise the claim is trivially true), and let us suppose to the contrary that both expressions are nonnegative. Then

$$\max_{i \in A, j \in A^c} \hat{\beta}_i \ominus V_{ij} = \max_{i \in A, j \in A^c} \hat{\beta}_j \ominus V_{ji} > \max_{i \in A, j \in A^c} \beta_j \ominus V_{ji} = \max_{i \in A, j \in A^c} \beta_i \ominus V_{ij} \geqq \max_{i \in A, j \in A^c} \hat{\beta}_i \ominus V_{ij},$$

which is a contradiction. The other two cases follow similarly, and so the claim is true. This shows that solutions to the order balance equations do not differ arbitrarily; specifically, all the arcs that separate A from A^c are transient.

Hence in particular, whenever we can show that

$$\beta_i \ominus V_{ij} \geqq 0 \quad \text{for all } i, j, \text{ with } i \neq j, \text{ and } V_{ij} < +\infty, \tag{19}$$

there can only exist *one* solution to the order balance equations for the given (ρ, V).

Now we show that this is indeed the case when ρ is large, which will prove the first part of the assertion (3) above. Specifically, suppose now that $\rho \geqq 2 \sum_{i,j:\, V_{ij} < +\infty} V_{ij}$.

Let $i_0 \in X$ be a state with $\beta_{i_0} = \rho$. For arbitrary $s \in X$, let $(i^* = i_0, i_1, \cdots, i_p = s)$ be a path from i^* to s such that $V_{i_{k-1}, i_k} < +\infty$ for $k = 1, \cdots, p$ and $i_k \neq i_m$ for $k \neq m$. Let $l(i) = \arg\min_j V_{ij}$. With $A = \{i_k\}$ and applying the Order Balance Theorem 2, it is easy to see that

$$\beta_{i_{k-1}} \ominus V_{i_{k-1}, i_k} \leqq \beta_{i_k} \ominus V_{i_k, l(i_k)} = \max_{j \neq i_k} (\beta_{i_k} \ominus V_{i_k, j}). \tag{20}$$

To prove that $\beta_s \geqq \max_{i,j:\, V_{ij}} V_{ij} < +\infty$, it is sufficient to show that for $k = 1, \cdots, p$, along the path from i^* to s,

$$\beta_{i_k} \geqq \beta_{i_0} - V_{i_0, i_1} + V_{i_1, l(i_1)} - V_{i_1, i_2} + V_{i_2, l(i_2)} - \cdots - V_{i_{k-1}, i_k} + V_{i_k, l(i_k)}, \tag{21}$$

since $\beta_{i_0} = \rho \geqq 2 \sum_{i,j:\, V_{ij} < +\infty} V_{ij}$.

We prove (21) by induction. For $k = 1$, from (20) we see that

$$\beta_{i_0} \ominus V_{i_0, i_1} \leqq \beta_{i_1} \ominus V_{i_1, l(i_1)}. \tag{22}$$

Clearly, the left-hand side of (22) is nonnegative, implying that the right-hand side is also nonnegative. Thus, we can replace "$\ominus$" with "$-$" giving

$$\beta_{i_1} \geqq \beta_{i_0} - V_{i_0, i_1} + V_{i_1, l(i_1)}. \tag{23}$$

Now assume (21) holds for $k-1$. From (20) we have

$$\beta_{i_{k-1}} \ominus V_{i_{k-1}, i_k} \leqq \beta_{i_k} \ominus V_{i_k, l(i_k)}. \tag{24}$$

The left-hand side of (24) is nonnegative and so

$$\begin{aligned}\beta_{i_k} &\geqq \beta_{i_{k-1}} - V_{i_{k-1},i_k} + V_{i_k,l(i_k)} \\ &\geqq \beta_{i_0} - V_{i_0,i_1} + V_{i_1,l(i_1)} - V_{i_1,i_2} + V_{i_2,l(i_2)} - \cdots - V_{i_{k-1},i_k} + V_{i_k,l(i_k)},\end{aligned}$$

which completes the induction proof. This proves (19), and therefore there exists a unique solution whenever ρ is large enough, which is the first half of assertion (3) above.

Moreover, for the large enough ρ specified earlier, due to (19), we have $\beta_i \ominus V_{ij} = \beta_i - V_{ij}$. Hence $\{\beta_i\}$ itself satisfies the modified balance equations. In fact, this solution is unique to the modified balance equations since, if λ is any other solution, then we can prove in a fashion similar to the above, that $\lambda_i \geqq V_{ij}$ for all $j \in N_i$, thus yielding that $\lambda_i \ominus V_{ij} = \lambda_i - V_{ij}$, which in turn proves that λ is yet another solution to the order balance equations, which is a contradiction.

Hence, at least for large enough values of ρ we have proved the existence of a unique solution to the modified balance equations. However, it is easy to see that if λ satisfies the modified equations for a given (ρ, V), then $\lambda - \delta$ satisfies the modified balance equations for $(\rho - \delta, V)$, thus proving the existence of a unique solution to the modified balance equations for *all* (ρ, V). This proves the assertion (2) as well as the second half of the assertion (3) above.

Now we turn to the proof of assertion (1) above. Let A be arbitrary, and let $\{\lambda_i\}$ be the solution of the modified balance equations, and define $\beta_i := \lambda_i \ominus 0$. Suppose

$$\max_{i \in A, j \in A^c} \lambda_i - V_{ij} < 0.$$

Then by (16) we also have

$$\max_{i \in A, j \in A^c} \lambda_j - V_{ji} < 0.$$

However, then for each $i \in A$ and $j \in A^c$,

$$\beta_i \leqq \lambda_i < V_{ij} \quad \text{and} \quad \beta_j \leqq \lambda_j < V_{ji}.$$

Hence,

$$\beta_i \ominus V_{ij} = -\infty \quad \text{and} \quad \beta_j \ominus V_{ji} = -\infty,$$

and so

$$\max_{i \in A, j \in A^c} \beta_i \ominus V_{ij} = \max_{i \in A, j \in A^c} \beta_j \ominus V_{ji},$$

thus satisfying the original order balance equations. If, however,

$$\max_{i \in A, j \in A^c} \lambda_i - V_{ij} = \delta \geqq 0,$$

then by (16)

$$\max_{i \in A, j \in A^c} \lambda_j - V_{ji} = \delta \geqq 0.$$

Suppose that $(i_1, j_1) \in A \times A^c$ and $(i_2, j_2) \in A^c \times A$ are such that

$$\lambda_{i_1} - V_{i_1,j_1} = \lambda_{j_2} - V_{j_2,i_2} = \delta.$$

Then since

$$\lambda_{i_1} = V_{i_1,j_1} + \delta \geqq 0 \quad \text{and} \quad \lambda_{j_2} = V_{j_2,i_2} + \delta \geqq 0$$

we have

$$\beta_{i_1} = \lambda_{i_1} \quad \text{and} \quad \beta_{j_2} = \lambda_{j_2},$$

and so

$$\beta_{i_1} - V_{i_1,j_1} = \beta_{j_2} - V_{j_2,i_2}.$$

Also, since $\lambda_k \geqq \beta_k$, we have

$$\begin{aligned}\max_{i\in A, j\in A^c} \beta_i \ominus V_{i,j} &\leqq \max_{i\in A, j\in A^c} \lambda_i \ominus V_{ij} \\ &\leqq \max_{i\in A, j\in A^c} \lambda_i - V_{ij} \\ &= \lambda_{i_1} - V_{i_1,j_1} \\ &= \beta_{i_1} - V_{i_1,j_1} \\ &= \beta_{i_1} \ominus V_{i_1,j_1}.\end{aligned}$$

Similarly, $\max_{i\in A, j\in A^c} \beta_j \ominus V_{ji} = \beta_{j_2} \ominus V_{j_2,i_2}$, and so

$$\max_{i\in A, j\in A^c} \beta_i \ominus V_{ij} = \max_{i\in A, j\in A^c} \beta_j \ominus V_{ji}.$$

This proves the assertion (1) and the theorem. □

Remark 1. It is interesting to note that the existence of a solution to the modified balance equations has been proved by relying on the existence of a solution to the order balance equations, which in turn is guaranteed by the probabilistic arguments of Theorem 2. A separate independent constructive proof of existence, which does not use probabilistic arguments, can be found in [8].

We now give an algorithm for determining the *unique* solution to the modified balance equations. An illustrative example is convenient.

Example 2. Let $\rho = 5$ and

$$V = [V_{ij}] = \begin{Bmatrix} \bigstar & 4 & 3 & 1 \\ 6 & \bigstar & 3 & 7 \\ 6 & 2 & \bigstar & 4 \\ 2 & 6 & 5 & \bigstar \end{Bmatrix}.$$

Our goal is to determine $\lambda = (\lambda_1, \cdots, \lambda_4)$, which satisfies (16), (17). We shall refer to $\lambda_i - V_{ij}$ as the λ-*flow* along the arc (i, j). Consider first the modified balance equation for the edge cut $A = \{i\}$,

$$\max_{j\neq i} \lambda_i - V_{ij} = \max_{j\neq i} \lambda_j - V_{ji}. \tag{25}$$

Observe that the left-hand side of (25) can be written as

$$\lambda_i - \min_{j\neq i} V_{ij},$$

and so the arc of *maximum* λ-*flow* out of $A = \{i\}$ is the arc $(i, l(i))$ where

$$l(i) = \arg\min_{j\neq i} V_{ij}.$$

(Note that $l(i)$ may not be unique.)

We now construct the directed graph $G_1 = (V_1, E_1)$, with $V_1 = \{\{1\}, \cdots, \{4\}\}$ and $(i, j) \in E_1$ if $j = l(i)$. See Fig. 1.

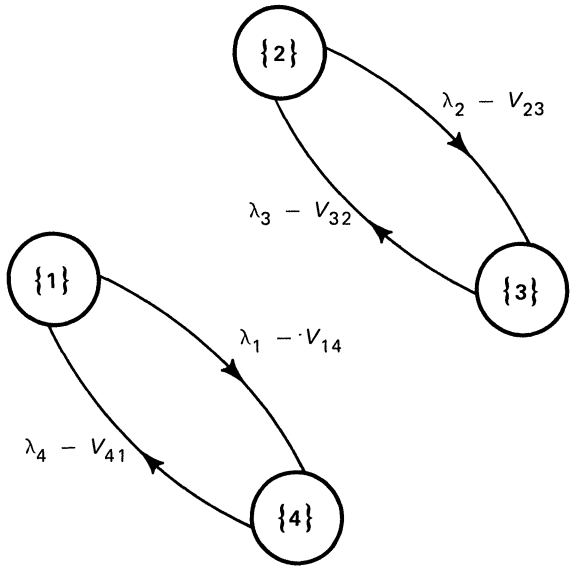

FIG. 1. *The graph G_1 of Example 2.*

Note that G_1 has two directed cycles $\{1\}\to\{4\}\to\{1\}$ and $\{2\}\to\{3\}\to\{2\}$. Let us examine the λ-flows on the directed cycle $\{1\}\to\{4\}\to\{1\}$. Since $\lambda_1 - V_{14}$ is the maximum λ-flow out of $\{1\}$, it is not smaller than any λ-flow into $\{1\}$, and so in particular

$$\lambda_1 - V_{14} \geqq \lambda_4 - V_{41}.$$

Also, $\lambda_4 - V_{41}$ is the maximum λ-flow out of $\{4\}$ and so

$$\lambda_4 - V_{41} \geqq \lambda_1 - V_{14}.$$

We thus observe that the λ-flows along the directed cycle $\{1\}\to\{4\}\to\{1\}$ are *equal*; that is,

$$\lambda_1 - V_{14} = \lambda_4 - V_{41},$$

and so

$$\lambda_1 - 1 = \lambda_4 - 2. \tag{26}$$

Thus, we have determined the *difference* between λ_1 and λ_4.

In exactly the same way, from the directed cycle $\{2\}\to\{3\}\to\{2\}$ we see that

$$\lambda_2 - 3 = \lambda_3 - 2, \tag{27}$$

thus determining the difference between λ_2 and λ_3.

At the next step of the algorithm, consider the modified balance equations for the edge cut (A, A^c) where $A = \{1, 4\}$ and $A^c = \{2, 3\}$. Observe that for $A = \{1, 4\}$, the left-hand side of the modified balance equation

$$\max_{i\in A, j\in A^c} \lambda_i - V_{ij} = \max_{i\in A, j\in A^c} \lambda_j - V_{ji} \tag{28}$$

can be written as

$$\max(\lambda_1 - V_{12}, \lambda_1 - V_{13}, \lambda_4 - V_{42}, \lambda_4 - V_{43});$$

that is,

$$\max(\lambda_1 - 4, \lambda_1 - 3, \lambda_4 - 6, \lambda_4 - 5).$$

We have previously determined that $\lambda_4 - \lambda_1 = 1$, and so the maximum is achieved by $\lambda_1 - V_{13} = \lambda_1 - 3$, and the arc of maximum λ-flow out of $\{1, 4\}$ is the arc $(1, 3)$.

In a similar fashion, examining the right-hand side of the modified balance equation (28), we determine that the maximum λ-flow out of $\{2, 3\}$ is achieved by $\lambda_3 - V_{34} = \lambda_3 - 4$, and so the arc of maximum λ-flow out of $\{2, 3\}$ is $(3, 4)$.

We now consider the directed graph $G_2 = (V_2, E_2)$, with $V_2 = \{\{1, 4\}, \{2, 3\}\}$ and $E_2 = \{(1, 3), (3, 4)\}$ shown in Fig. 2. Note that E_2 is the set of the arcs of maximum λ-flow out of the edge cuts in V_2.

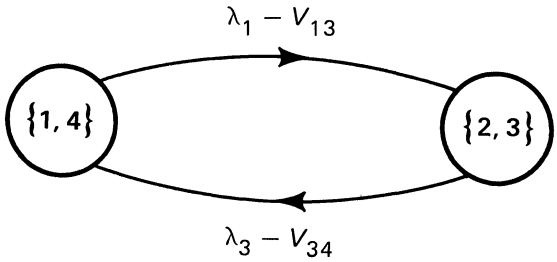

FIG. 2. *The graph G_2 of Example 2.*

Observe that G_2 has a directed cycle $\{1, 4\} \to \{2, 3\} \to \{1, 4\}$. Now note that $\lambda_1 - V_{13}$ is the maximum λ-flow out of $\{1, 4\}$ and $\lambda_3 - V_{34}$ is the maximum λ-flow out of $\{2, 3\}$ and so

$$\lambda_1 - V_{13} = \lambda_3 - V_{34};$$

that is,

$$\lambda_1 - 3 = \lambda_3 - 4. \tag{29}$$

Combining (26), (27), and (29), we obtain

$$\lambda_1 - 3 = \lambda_2 - 5 = \lambda_3 - 4 = \lambda_4 - 4. \tag{30}$$

We now know the *pairwise differences* between *all* of the λ_i's, and so we do not need to consider any additional edge cuts. To fix the values of $\{\lambda_i\}$, we use the value of ρ to give

$$\max_{i \in X} \lambda_i = \rho = 5.$$

Since, from (30), λ_2 is the largest, we set $\lambda_2 = 5$. We thus obtain the solution to the modified balance equations:

$$\lambda_1 = 3, \quad \lambda_2 = 5, \quad \lambda_3 = \lambda_4 = 4.$$

The principal idea used to solve the modified balance equations in Example 2 is summarized in the following lemma.

LEMMA 2. (1) *Given $A \subseteq X$ for which we know the pairwise differences between all the λ_i's for states in A, we can determine the arc of maximum λ-flow out of A (without knowing the λ_i's themselves).*

(2) *Let $A_1, A_2, \cdots, A_p$ be a partition of X and suppose for each A_k we know all the pairwise differences between the λ_i's for all states in A_k. Let (i_k, j_k) denote the arc of maximum λ-flow out of A_k. Construct the directed graph $G = (V, E)$, with $V = \{A_1, \cdots, A_p\}$ and $E = \{(i_1, j_1), \cdots, (i_p, j_p)\}$. There exists a directed cycle on G. If*

$\{A_{n_1}, \cdots, A_{n_m}\}$ is the list of vertices, in order, along the directed cycle, then the λ-flow on the directed cycle is constant; that is,

$$\lambda_{i_{n_1}} - V_{i_{n_1}, j_{n_1}} = \cdots = \lambda_{i_{n_m}} - V_{i_{n_m}, j_{n_m}},$$

and we can determine the pairwise differences between the values of the λ_i's for all the states in $\bigcup_{k=1}^{m} A_{n_k}$.

Proof. (1) Without loss of generality, suppose A is the set of states $\{1, 2, \cdots, r\}$. Let $\alpha_i := \lambda_1 - \lambda_i$. (We know the α_i's.) Then

$$\max_{i \in A, j \in A^c} \lambda_i - V_{ij} = \max_{i \in A, j \in A^c} \lambda_1 - \alpha_i - V_{ij} = \lambda_1 - \min_{i \in A, j \in A^c} (\alpha_i + V_{ij}).$$

Thus, the arc

$$(i^*, j^*) := \arg \min_{i \in A, j \in A^c} (\alpha_i + V_{ij})$$

is an arc of maximum λ-flow out of A.

(2) The out-degree of each vertex of G is at least one, and so from elementary graph theory it follows that G has a directed cycle. Suppose

$$A_{n_1} \to A_{n_2} \to \cdots \to A_{n_m} \to A_{n_1}$$

is such a directed cycle. Then we have the situation shown in Fig. 3. Now (i_{n_k}, j_{n_k}) is the arc of maximum λ-flow *out* of A_{n_k}, and so the λ-flow on this arc is not less than the λ-flow of any arc *into* A_{n_k}. In particular,

$$\lambda_{i_{n_k}} - V_{i_{n_k}, j_{n_k}} \geqq \lambda_{i_{n_{k-1}}, j_{n_{k-1}}} - V_{i_{n_{k-1}}, j_{n_{k-1}}} \quad \text{for } k = 1, \cdots, m,$$

where, for convenience, we implicitly identify i_{n_0} with i_{n_m} and j_{n_0} with j_{n_m}. Thus,

$$\begin{aligned} \lambda_{i_{n_m}} - V_{i_{n_m}, j_{n_m}} &\geqq \lambda_{i_{n_{m-1}}} - V_{i_{n_{m-1}}, j_{n_{m-1}}} \\ &\geqq \lambda_{i_{n_{m-2}}} - V_{i_{n_{m-2}}, j_{n_{m-2}}} \\ &\vdots \\ &\geqq \lambda_{i_{n_1}} - V_{i_{n_1}, j_{n_1}} \\ &\geqq \lambda_{i_{n_m}} - V_{i_{n_m}, j_{n_m}}. \end{aligned}$$

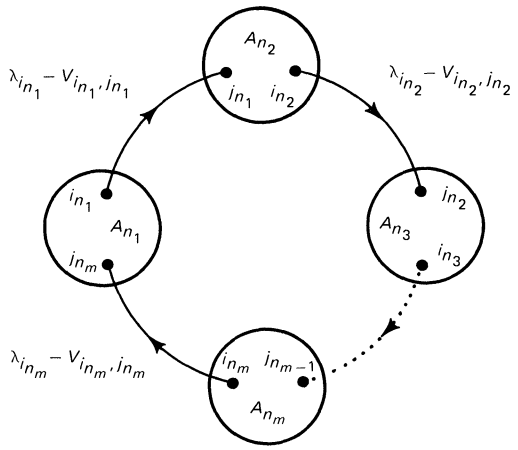

FIG. 3. *A directed cycle of maximum λ-flows in Lemma* 3.

Therefore, the λ-flow on the directed cycle is a constant:

$$\lambda_{i_{n_1}} - V_{i_{n_1},j_{n_1}} = \lambda_{i_{n_2}} - V_{i_{n_2},j_{n_2}} = \cdots = \lambda_{i_{n_m}} - V_{i_{n_m},j_{n_m}}. \tag{31}$$

For each A_{n_k} in the directed cycle, we know the pairwise differences between the λ_i's for states in A_{n_k}. Using (31) we can now easily determine the pairwise differences between all the λ_i's for states in $\cup_{k=1}^{m} A_{n_k}$. □

The algorithm for solving the modified balance equations is outlined below.

ALGORITHM TO SOLVE MODIFIED BALANCE EQUATIONS.

Step 1. Set $A_i^1 = \{i\}$ for $i = 1, \cdots, |X|$. We call the A_i^k's *coalitions at step k.* Note that for every i, the pairwise differences between the λ-values for all states in A_i^1 are (trivially) known. Set $A^1 := \{A_1^1, A_2^1, \cdots, A_{|X|}^1\}$. Let $N(1) = |A^1| =:$ the number of elements in the set $A^1 =$ number of coalitions at Step 1.

Step k. Given $A^k := \{A_1^k, A_2^k, \cdots, A_{N(k)}^k\}$, where for each $A_j^k \in A^k$ the pairwise differences between all of the λ_i's for i's in A_j^k are known, construct A^{k+1} as follows. Using Lemma 2, identify the directed cycles in the graph. (There exists at least one directed cycle.) The elements of A^{k+1} consist of the directed cycles identified in the graph, and those $A_j^k \in A^k$ that are not in any directed cycle. (More precisely, if $\{A_{n_1}^k, A_{n_2}^k, \cdots, A_{n_m}^k\}$ is a directed cycle, then $\cup_{i=1}^{m} A_{n_i}$ is an element of A^{k+1}.) Note that for every $A_j^{k+1} \in A^{k+1}$, the pairwise differences between all of the λ_i's for i's in A_j^{k+1} are known. Furthermore, if $N(k) := |A^k|$, then $N(k+1) < N(k)$.

Last Step. Stop when $N(k) = 1$. Note that the pairwise differences between *all* λ_i's are known, and the λ satisfying the modified balance equations can be obtained by a translation by using the given value of ρ. □

4. An algorithm to obtain all solutions of the order balance equations. We now characterize *all* solutions to the order balance equations, and describe an algorithm for generating *all* these solutions. To do so we will use the coalitions $\{A_i^k\}$ generated by the algorithm of the preceding section. Let us call $\lambda_i - V_{ij}$ and $\beta_{ij} = \beta_i \ominus V_{ij}$ as the *λ-flow* and *β-flow*, respectively, along the arc (i, j).

LEMMA 3. (1) *If (i, j) is an arc of maximum λ-flow out of A_l^k, then it is also an arc of maximum β-flow out of A_l^k.*

(2) *If $\{A_1^k, \cdots, A_p^k\}$ is a directed cycle obtained at step k, then the β-flow along the directed cycle is a constant.*

(3) *If the β-flow along the directed cycle $\{A_1^k, \cdots, A_p^k\}$ obtained at step k is $-\infty$, then the β-flow along any directed cycle obtained at step $n > k$ containing $A_l^n = \cup_{i=1}^{p} A_i^k$ as a node, is also $-\infty$.*

(4) *If the β-flow along the directed cycle $\{A_1^k, \cdots, A_p^k\}$ obtained at step k is $\geqq 0$, then for every $i, j \in A_l^{k+1} := \cup_{m=1}^{p} A_m^k$ there exists a path $(i = i_0, i_1, \cdots, i_q = j)$ such that $i_m \in A_l^{k+1}$ and $\beta_{i_m, i_{m+1}} \geqq 0$ for $0 \leqq m \leqq q-1$.*

Proof. We will first prove (1)–(3) by induction. Consider $k = 1$. Since A_l^k is then just a singleton, say $A_l^k = \{l\}$, an arc (l, m) of maximum λ-flow out of $\{l\}$ is just one for which $V_{lm} = \min_n V_{ln}$. Clearly this is also an arc for which $\beta_l \ominus V_{lm} = \min_n \beta_l \ominus V_{ln}$. Now suppose that $\{A_1^k, \cdots, A_p^k\}$ is a directed cycle of such maximum flows. Then an application of the Order Balance Theorem to each A_l^k shows that $\beta_{12} = \beta_{23} = \cdots = \beta_{p1}$. Suppose now that $\beta_{12} = \beta_{23} = \cdots = \beta_{p1} = -\infty$. Then if (l, m) is an arc of maximum β-flow out of $\cup_{i=1}^{p} A_i^k$, clearly $\beta_{lm} \leqq \beta_{l,l+1} = -\infty$. Thus the assertion is true for $k = 1$.

Now suppose that the assertion is true for $1, 2, \cdots, k-1$. Consider a coalition A_l^k. If the β-flow along some directed cycle $\{A_1^n, \cdots, A_q^n\}$ at some step $n < k$ with $A_l^k = \cup_{i=1}^{q} A_i^n$ was $-\infty$, then clearly the maximum β-flow out of A_l^k is $-\infty$, and so *any*

arc out of A_l^k is an arc of maximum β-flow. On the other hand if the β-flow along the directed cycle $\{A_1^n, \cdots, A_q^n\}$ is $\geqq 0$, then the differences between the β_i's for states $i \in A_l^k$ are the *same* as the differences between the λ_i's, i.e.,

$$\beta_i - \beta_j = \lambda_i - \lambda_j \quad \text{for all } i, j \in A_l^k, \tag{32}$$

and so the arc of maximum λ-flow out from A_l^k is also an arc of maximum β-flow out from A_l^k. Moreover, if $\{A_1^k, \cdots, A_p^k\}$ is a directed cycle at step k, then an application of the Order Balance Theorem to each A_l^k shows that the β-flow along the directed cycle is a constant. Finally, if this β-flow is $-\infty$, suppose that (r, m) is a maximum flow arc out of $\cup_{l=1}^p A_l^k$. Suppose that $r \in A_l^k$. Then clearly $\max_{i \in A_l^k, j \in A_l^{k^c}} \beta_{ij} \geqq \beta_{rm}$ and so $\beta_{rm} = -\infty$. This completes the induction and the proof.

Finally, to see (4), note first that from (1), (2), and (3), the β-flow along any directed cycles contained within A_l^{k+1} is $\geqq 0$. Since A_l^{k+1} is formed as the union of such directed cycles, the result follows. □

Motivated by (3) and (4) above, we introduce the following definition.

DEFINITION 4. We shall say that i is *recurrently connected* to j if there exists a path $(i = i_0, i_1, \cdots, i_q = j)$ with $\beta_{i_m, i_{m+1}} \geqq 0$ for $0 \leqq m \leqq q-1$.

We shall say that a set $A \subseteq X$ is a *recurrently connected set* if for every $i, j \in A$ and $k \in A^c$, i is recurrently connected to j but not to k.

From Lemma 3 it follows that recurrently connected sets are precisely those A_l^k's for which the β-flow out of A_l^k is $-\infty$, while the β-flows along the directed cycles contained within A_l^k are $\geqq 0$. Note also that the recurrently connected sets form a *partition* of X.

We now proceed to determine which sets are possible candidates for being recurrently connected sets. Consider a typical candidate A_l^{k+1}. Let $\mathscr{F}$ denote the β-flow on the cycle $\{A_1^k, \cdots, A_p^k\}$, where $A_l^{k+1} = \cup_{i=1}^p A_i^k$. Then if (i_m, j_m) is the arc of maximum flow out of A_m^k (and, by construction, into $A_{(m+1) \bmod p}^k$), we must have

$$\mathscr{F} = \beta_{i_1} - V_{i_1, j_1} = \beta_{i_2} - V_{i_2, j_2} = \cdots = \beta_{i_p} - V_{i_p, j_p} \geqq 0,$$

$$\max_{i \in A_l^{k+1}, j \notin A_l^{k+1}} \beta_i - V_{ij} < 0, \quad \max_{i \in A_l^{k+1}} \beta_i \leqq \rho.$$

We will now attempt to determine whether there exist $\{\beta_i : i \in A_l^{k+1}\}$ that satisfy these conditions. Note that if this is not feasible, then A_l^{k+1} cannot be a recurrently connected set.

Let (x, y) denote the arc of maximum β-flow out of A_l^{k+1}. Then $\beta_x < V_{xy}$. Fix m to be an arbitrarily chosen state from A_l^{k+1}. Then for every state $h \in A_l^{k+1}$ we know the value of $(\beta_h - \beta_m)$ from Lemma 3 above. Let us define

$$\zeta_h := \beta_h - \beta_m.$$

Then

$$\begin{aligned} \mathscr{F} &= \beta_{i_1} - V_{i_1, j_1} \\ &= \beta_m + \zeta_{i_1} - V_{i_1, j_1} \\ &= \beta_x - \zeta_x + \zeta_{i_1} - V_{i_1, j_1} \\ &< V_{xy} - \zeta_x + \zeta_{i_1} - V_{i_1, j_1} =: M_1, \end{aligned}$$

giving an upper bound on $\mathscr{F}$.

We must also satisfy the constraint $\max_{i \in A_l^{k+1}} \beta_i \leqq \rho$, and so let

$$\theta := \arg \max_{i \in A_l^{k+1}} \zeta_i.$$

Then it is clear that $\beta_\theta \geqq \max_{i \in A_l^{k+1}} \beta_i$. Thus,

$$\begin{aligned}\rho &\geqq \beta_\theta \\ &= \beta_m + \zeta_\theta \\ &= \beta_{i_1} - \zeta_{i_1} + \zeta_\theta \\ &= \beta_{i_1} - V_{i_1,j_1} + V_{i_1,j_1} - \zeta_{i_1} + \zeta_\theta \\ &= \mathscr{F} + V_{i_1,j_1} - \zeta_{i_1} + \zeta_\theta,\end{aligned}$$

and so

$$\mathscr{F} \leqq \rho - V_{i_1,j_1} + \zeta_{i_1} - \zeta_\theta =: M_2,$$

giving yet another upper bound on $\mathscr{F}$. (Note. If $A_l^{k+1} = \{i\}$, then $M_1 = \min_j V_{ij}$ and $M_2 = \rho$.)

Any choice of $\mathscr{F}$ from the interval

$$\Omega(A_l^{k+1}) := [0, M_1) \cap [0, M_2]$$

will allow assignments for the recurrence orders of states in A_l^{k+1} consistent with the assumption that the coalition A_l^{k+1} is a recurrently connected set. If $\Omega(A_l^{k+1}) = \varnothing$ then then there is no assignment, and so A_l^{k+1} is not a recurrently connected set.

We still need to determine the set of all recurrently connected sets. To do this we construct a *rooted tree* having the coalitions produced by the general procedure as nodes, and having a directed edge from coalition A_p^{k+1} to A_r^k if $A_p^{k+1} \supseteq A_r^{(k)}$. Hence, the root of the tree is X, and its leaves are the singleton sets $\{1\}, \{2\}, \cdots, \{n\}$. Let D_i be the set of the leaves of the tree that are *descendants* of the node i in the rooted tree.

We say that a set Ξ of nodes is a *proper cover* if

$$\bigcup_{A \in \Xi} D_A = X$$

and

$$D_A \cap D_{A'} = \varnothing \quad \text{for } A \neq A'.$$

Now the algorithm to determine all the solutions of (15), (8) proceeds as follows. Let a set $\Xi := \{A_1, A_2, \cdots, A_k\}$ be a proper cover. Now we will determine whether Ξ can be a set of *all* recurrently connected sets, as follows. First we determine $\Omega(A_j)$ for every $A_j \in \Xi$. (Note that if we guess X to be a recurrently connected set, then $\Omega(X) = [0, M_2]$, since the M_1 upper bound is $+\infty$ because there is no maximal flow out of X. Also, if we guess the singleton $\{i\}$ to be a recurrently connected set, then $\Omega(\{i\}) = -\infty \cup ([0, M_1) \cap [0, M_2])$. If *any* of the $\Omega(A_j)$'s is empty, then the guess Ξ is *not* a feasible set of recurrently connected sets. If *every* $\Omega(A_j)$ is nonempty, then let $\bar{\mathscr{F}}_j := \sup \Omega(A_j)$. If this "sup" is not attained, then we cannot assign ρ to any state in A_j. If this "sup" is attained, then we determine for each such A_j whether, with the choice of $\bar{\mathscr{F}}_j$, there is a state $i_j \in A_j$ with $\beta_{i_j} = \rho$. If no such state exists for *any* A_j, then again Ξ is *not* a feasible set of recurrently connected sets. Finally, if there exist such A_j's then let $\mathscr{A}(\Xi)$ be the set of all such A_j's. Now, the set of all solutions corresponding to Ξ is obtained by picking, in turn, an A_j from $\mathscr{A}(\Xi)$, fixing its flow as $\bar{\mathscr{F}}_j$ and choosing all other $\mathscr{F}_j$'s arbitrarily from the $\Omega(A_j)$'s. By checking *every* proper cover Ξ, we thus determine all solutions to the order balance equations, as the following theorem shows.

THEOREM 4. *All solutions to the order balance equations can be generated by using the method described above.*

Proof. Suppose β satisfies the order balance equations. Then for this solution determine the set Ξ of recurrently connected sets. This set must be a proper cover. For this set Ξ, there must be some A_j with corresponding β-flow equal $\bar{\mathcal{F}}_j$. Now determine the β-flows on the recurrently connected sets. We generate this solution β when we choose Ξ as the set of recurrently connected sets, and A_j as the coalition with maximum flow equal to $\bar{\mathcal{F}}_j$, and assign the correct β-flows on the other recurrently connected sets. □

This algorithm takes an exponential in $|X|$ number of steps, due to the necessity of checking all proper covers. However, the complexity issue is not the primary concern here, since the problem of asymptotic analysis of the stochastic process is not a priori known to be a problem resolvable by a finite algorithm.

We illustrate the procedure for determining all solutions to the order balance equations.

Example 3. We construct all solutions to the order balance equations for Example 2 when $\rho = 4$. See Fig. 4 for the rooted tree. We check the proper covers:

(1) $\Xi = \{X\}$: $\Omega(X)$ is empty, so X cannot be a recurrently connected set.

(2) $\Xi = \{\{1, 4\}, \{2, 3\}\}$: Using the method described above we obtain

$$\beta_1 = \alpha, \quad \beta_2 = 4, \quad \beta_3 = 3, \quad \beta_4 = 1 + \alpha$$

where $1 \leqq \alpha < 3$.

(3) $\Xi = \{\{1, 4\}, \{2\}, \{3\}\}$: $\max_{i \in X} \beta_i < 4$, a contradiction.

(4) $\Xi = \{\{1\}, \{4\}, \{2, 3\}\}$:

$$\beta_1 = \gamma, \quad \beta_2 = 4, \quad \beta_3 = 3, \quad \beta_4 = \theta$$

where $\gamma = -\infty$ or $0 \leqq \gamma < 1$, and $\theta = -\infty$ or $0 \leqq \theta < 2$.

(5) $\Xi = \{\{1\}, \{2\}, \{3\}, \{4\}\}$: $\max_{i \in X} \beta_i < 4$, and so $\{\{1\}, \{2\}, \{3\}, \{4\}\}$ is not a set of recurrently connected sets.

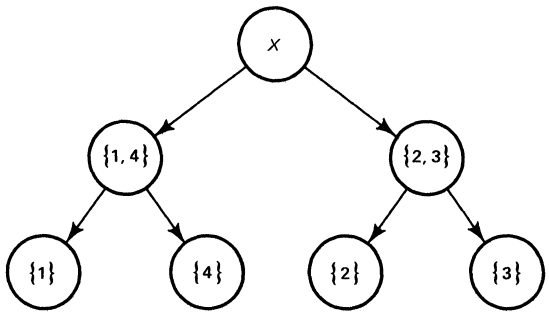

FIG 4. *The rooted tree of Example* 3.

We have checked all proper covers. Hence the set of all solutions is $\{(\alpha, 4, 3, 1+\alpha)$: $1 \leqq \alpha < 3\} \cup \{(\gamma, 4, 3, \theta)$: $\gamma = -\infty$ or $0 \leqq \gamma < 1$ and $\theta = -\infty$ or $0 \leqq \theta < 2\}$.

How can nonunique solutions to the order balance equations arise, and what is the implication of such nonuniqueness? First let us consider the case where a unique solution exists. Since such a solution is uniquely determined by the algorithm, it is clear that the recurrence orders of the states, and thus the rates of convergence of the transition probabilities, depend only on the V_{ij}'s in the transition probabilities $p_{ij}(t) = c_{ij}\varepsilon(t)^{V_{ij}}$, and *not* on the proportionality constants $\{c_{ij}\}$. However, in the case of *nonunique* solutions, the following example shows that the recurrence orders may even depend on the *proportionality constants* $\{c_{ij}\}$.

Example 4. Let $X = \{1, 2, 3\}$ and $V_{ij} = \max\{0, j - i\}$. Let $c_{13} = c_{23} = 1$, $c_{31} = 1 - \alpha$, and $c_{32} = \alpha$, where $\alpha \in (0, 1)$. Set $c_{ij} = 0$ for all other i, j. See Fig. 5. Let the cooling schedule be $\varepsilon(t) = 1/t$. Then the *complete* set of order balance equations obtained by using *all* edge cuts is:

$$\beta_2 \ominus V_{23} = \beta_3 \ominus V_{32}, \qquad \beta_3 \ominus V_{31} = \beta_1 \ominus V_{13},$$

$$\max(\beta_2 \ominus V_{23}, \beta_1 \ominus V_{13}) = \max(\beta_3 \ominus V_{32}, \beta_3 \ominus V_{31}),$$

with the maximum given by,

$$\max_{i \in X} \beta_i = 1.$$

The assignments

$$\beta_1 = 1, \quad \beta_2 = \gamma, \quad \beta_3 = -\infty$$

satisfy the order balance equations for *every* $\gamma \in \{-\infty\} \cup [0, 1)$. Thus any value of $\beta_2 < 1$ gives a solution of the order balance equations.

However, a calculation that can be found in [8] shows that the correct order of recurrence of state 2 is

$$\beta_2 = \alpha.$$

Thus, the order of recurrence, and the rate of convergence of the probability $\Pr(x(t) = 2)$ to zero, depends on the proportionality constant $c_{32} = \alpha$ involved.

Based on the above results, we obtain the following property of the orders of recurrence of the states in a recurrently connected set.

LEMMA 4. *Consider a recurrently connected set A.*

(1) *If* $\beta_i \in \mathcal{R}$ *for some* $i \in A$, *then* $\beta_j \in \mathcal{R}$ *for all* $j \in A$.

(2) *If for some* $i \in A$, $\beta_i = p_i^-$ *for some* $p_i \in \mathcal{R}$, *then for every* $j \in A$, $\beta_j = p_j^-$ *for some* $p_j \in \mathcal{R}$.

Proof. The proof follows immediately from (32). □

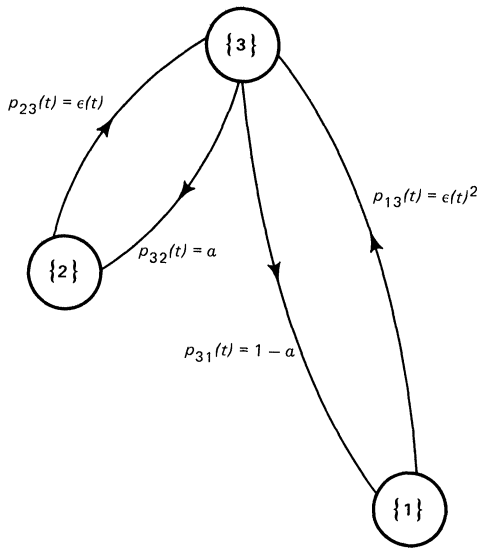

FIG. 5. *The Markov process of Example* 4.

Thus all recurrence orders in a recurrently connected set are of the same type, i.e., either they are all real numbers p_i, or they are all of the type p_i^-, or they are all $-\infty$ (see Definition 1).

This gives us the following lemma, which completes the proof of Theorem 1.

LEMMA 5. *Suppose the rate of cooling is* $\rho = p \in \mathcal{R}$, *with* $p > 0$, *i.e., the maximum is achieved in Definition* 3. *If there is a state* $i \in X$ *for which* $\beta_i = p^-$, *then* $\lim_{t\to\infty} \Pr(x(t) = i) = 0$.

Proof. Suppose A is the recurrently connected class to which i belongs. Since all arcs *between* recurrently connected sets are transient, it follows from the Borel-Cantelli Lemma that along almost every sample path ω there can only be a *finite* number of transitions between different recurrently connected sets. Hence for almost every ω, $\{x(t, \omega)\}$ converges to some recurrently connected set. Hence the limit $\lim_{t\to\infty} \Pr(x(t) \in A)$ exists. Now we show that this limit is zero. Suppose not, i.e., suppose $\lim_{t\to\infty} \sum_{j\in A} \pi_j(t) = \delta > 0$. Then it follows that $\sum_{t=0}^{\infty} \varepsilon(t)^p \sum_{j\in A} \pi_j(t) = +\infty$. Hence for some $j \in A$, $\beta_j = p$. But then by Lemma 4, $\beta_i \in \mathcal{R}$, which gives a contradiction. □

5. Weak reversibility and simulated annealing. We now turn our attention to the special class of Markov chains arising from the method of *optimization by simulated annealing*. Recall that the Markov chains in this class satisfy (1)-(6) with the special choice of

$$V_{ij} := \max\{0, W_j - W_i\}.$$

In [7] it was shown that under the "symmetric neighborhood" assumption, $c_{ij} > 0$ if and only if $c_{ji} > 0$, the orders of recurrence satisfy the following *detailed order balance*:

$$\beta_{ij} = \beta_{ji} \quad \text{for every } i, j \in X.$$

It is easy to see that the detailed order balance above is equivalent to the sum of the order of recurrence of a state and its cost being constant on recurrently connected sets.

In this section we will show that this constancy property of the sum of the recurrence order and cost on recurrently connected sets continues to hold under the much weaker assumption of "weak reversibility" introduced by Hajek in [1].

DEFINITION 5. A state i is said to be reachable from state j if there is a sequence of states $j = i_0, i_1, \cdots, i_p = i$ such that $c_{i_k, i_{k+1}} > 0$ for $0 \leqq k \leqq p-1$.

DEFINITION 6. A state i is reachable at height H from j if there is a path from j to i as in Definition 5 for which $W_{i_k} \leqq H$ for $0 \leqq k \leqq p$.

ASSUMPTION 1(Weak Reversibility). For any real number H and any two states i and j, i is reachable at height H from j if and only if j is reachable at height H from i.

In what follows we assume weak reversibility.

THEOREM 5 (The Potential Theorem). *Under Assumption* 1, *for every recurrently connected set* A *there exists a constant* $\alpha(A)$ *such that* $\beta_i + W_i = \alpha(A)$ *for every* $i \in A$.

Proof. We fix our attention on a particular recurrently connected set A. Assume to the contrary that A can be partitioned into *equipotential* sets $C_1, C_2, \cdots, C_r$ such that $\beta_i + W_i = \alpha(C_k)$ for every $i \in C_k$, where the $\alpha(C_k)$'s are distinct constants. We will show that there is only one equipotential set, namely, A.

For each equipotential set C_i, determine an arc of maximum β-flow out of the set. From Lemma 2, there exists a directed cycle of these equipotential sets, and the β-flow along the directed cycle is constant. Moreover, from Lemma 3, since A is a recurrently connected set, these β-flows are all nonnegative. Without loss of generality, label the sets along the directed cycle $C_1, C_2, \cdots, C_p$ such that the constant $\alpha(C_1)$ associated with the set C_1 is smallest. Let (i_s, j_s) be the arc of maximum β-flow out

of the set C_s. By construction, $i_s \in C_s$ and $j_s \in C_{(1+s) \bmod p}$ and

$$\beta_{i_1,j_1} = \beta_{i_2,j_2} = \cdots = \beta_{i_p,j_p} \geqq 0.$$

Knowing that $\beta_{i_1,j_1} \geqq 0$ we consider the two cases: (1) $W_{j_1} \geqq W_{i_1}$; or (2) $W_{j_1} < W_{i_1}$.

If case (1) is true then since j_1 is reachable at height W_{j_1} from i_1, by the weak reversibility assumption there exists a path from j_1 back to i_1 that does not go through any states with costs larger than W_{j_1}. Let (k, l) be the particular arc of that path that exits C_2. Note that

$$\begin{aligned}\beta_{i_1,j_1} &= \beta_{i_2,j_2} \\ &\geqq \beta_{kl},\end{aligned}$$

because β_{i_2,j_2} is the arc of maximum β-flow out of C_2. If $\beta_{kl} \geqq 0$ then $\beta_{kl} = \beta_k + W_k - W_l$. If $\beta_{kl} < 0$ then $\beta_k + W_k - W_l < 0$. In either case, since $\beta_{i_1,j_1} \geqq 0$, we have that

$$\beta_{i_1,j_1} = \beta_{i_1} + W_{i_1} - W_{j_1} \geqq \beta_k + W_k - W_l.$$

Now by the weak reversibility assumption, $W_{j_1} \geqq W_l$, and so

$$\beta_{i_1} + W_{i_1} \geqq \beta_k + W_k;$$

that is,

$$\alpha(C_1) \geqq \alpha(C_2),$$

which is a contradiction.

If case (2) is true, then there is a path from j_1 to i_1 that does not pass through any states with costs larger than i_1. Again, identify the particular arc of that path that exits C_2 as (k, l). Note that

$$\begin{aligned}\beta_{i_1} &= \beta_{i_1,j_1} \\ &= \beta_{i_2,j_2} \\ &\geqq \beta_{kl}.\end{aligned}$$

Using similar arguments as in case (1), since $\beta_{i_1} \geqq 0$ we have $\beta_{i_1} \geqq \beta_k + W_k - W_l$. Now by the weak reversibility assumption $W_{i_1} \geqq W_l$, and so $\alpha(C_1) \geqq \alpha(C_2)$, which is again a contradiction.

Hence there is only one equipotential set, A. □

Since $W_i + \beta_i = \alpha(A)$ for all $i \in A$, where A is a recurrently connected set, we obtain the following necessary and sufficient condition for simulated annealing to hit a *global minimum* with probability one from all states $i \in X$.

Let $M := \{i \in X: W_i \leqq W_j \text{ for all } j \in X\}$ be the set of global minima. We now have the following definition due to [1].

DEFINITION 7. Let d^* be the *smallest* number with the property that for every $i \in X$ there exists a path $(i = i_0, \cdots, i_p)$ with $c_{i_k,i_{k+1}} > 0$ for $0 \leqq k \leqq p-1$ and ending in a minimizer $i_p \in M$ such that

$$W_{i_k} - W_i \leqq d^* \quad \text{for } k = 1, \cdots, p.$$

We shall call d^* the *depth* of the minimization problem.

THEOREM 6 (Necessary and Sufficient Condition to Hit Global Minimum With Probability One). *Suppose that weak reversibility holds.*

(1) *If* $\sum_{t=1}^{\infty} \varepsilon(t)^{d^*} = +\infty$, *then for every initial condition* $x(0) \in X$,

$$\limsup_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \Pr(x(t) \in M) = 1,$$

and the global minimum is hit with probability one.

(2) *If* $\sum_{t=1}^{\infty} \varepsilon(t)^{d^*} < +\infty$, *then there exists an initial condition* $x(0) \in X$ *for which*

$$\Pr(x(t) \in M^c \text{ for all } t \geqq 1) > 0.$$

Proof. The proof is the same as in Theorem (4.6) in [7] except that if $(i = i_0, \cdots, i_p = j)$ is a path from i to j with $c_{i_k, i_{k+1}} > 0$ and $W_{i_k} - W_i \leqq \gamma$ for $1 \leqq k \leqq p$, then instead of using the reversed path $(j = i_p, \cdots, i_0 = i)$ given by the assumption of symmetric neighborhoods, we use the path $(j = l_0, \cdots, l_q = i)$ with $c_{l_k, l_{k+1}} > 0$ and $W_{l_k} - W_i \leqq \gamma$ for $1 \leqq k \leqq q$, guaranteed by the weak reversibility assumption. □

The same condition $\sum_{t=1}^{\infty} \varepsilon(t)^{d^*} = \infty$ has been shown earlier by Hajek [1] to be necessary and sufficient for $\lim_{t \to \infty} \Pr(x(t) \in M) = 1$, i.e., for convergence *in probability.* Thus while result (1) above is weaker than his, since it involves Cesaro as opposed to regular convergence, the result (2) is a stronger sample path result.

The above result has been proved earlier in [7] under the stronger assumption of *symmetric neighborhoods*, $c_{ij} > 0 \Leftrightarrow c_{ji} > 0$. Moreover, under this assumption Connors and Kumar [7] have proved a *detailed balance* result that we can obtain as a corollary of Theorem 5, as we show below.

COROLLARY 1 (Detailed Balance). *Under the symmetric neighborhood assumption,*

$$\beta_{ij} = \beta_{ji} \quad \text{for every } i, j \in X.$$

Proof. If i and j are not neighbors, then $\beta_{ij} = \beta_{ji} = -\infty$.

If i and j are neighbors and $i \in R$ and $j \in T$, where R is the set of recurrent states and T is the set of transient states, then

$$\beta_{jk} = -\infty \quad \text{for all } k$$

and so

$$-\infty = \max_{k \neq j} \beta_{jk} = \max_{k \neq j} \beta_{kj} \geqq \beta_{ij},$$

showing that $\beta_{ij} = \beta_{ji} = -\infty$. A similar argument holds if $i \in T$ and $j \in R$.

Finally, if i and j are neighbors and $i, j \in R$, without loss of generality let us assume that $W_i \geqq W_j$. Then $\beta_{ij} = \beta_i \geqq 0$, and so i and j belong to a common recurrently connected set. Hence by Theorem 5, $\beta_i + W_i = \beta_j + W_j$. Since $\beta_{ij} = \beta_i$ and $\beta_{ji} = \beta_j + W_j - W_i$, it follows that $\beta_{ij} = \beta_{ji}$. □

Note that by the above results, if the order of recurrence of even *one* state in a connected component is known, then the orders of recurrence for *all* the states belonging to the connected component are determined. However, as Example 4 shows, it is not always possible to determine the order of recurrence of even one state in a connected component from the order balance equations alone. In that example, the connected components of recurrent states are the sets $\{1\}$ and $\{2\}$, and the detailed balance equations do not determine the order of recurrence β_2 of the *single* state in the connected component $\{2\}$. The reason for this inadequacy, as mentioned earlier in Example 4, is that the orders of recurrence do depend on the proportionality constants c_{ij} involved in the transition probabilities. In any case, the β-flows do satisfy Corollary 1.

6. Conclusions. The notion of order of recurrence provides a novel approach for analyzing the class of Markov chains whose transition probabilities are proportional to powers of a time-varying parameter $\varepsilon(t)$. These recurrence orders satisfy a set of balance equations, and the Markov chain converges in a Cesaro sense to the set of states with the largest recurrence orders. We have given an algorithm for generating a solution to the order balance equations and have also provided a method for characterizing all solutions to these equations. The algebraic methods presented in this

paper for solving the order balance equations are not always sufficient for determining the recurrence orders. In some situations where nonunique solutions exist, the orders of recurrence can depend on the proportionality constants involved in the transition probabilities, and not just on their orders of magnitude. This problem remains an open issue. The method of optimization by simulated annealing falls within the framework of this class of Markov chains. We have shown that if the Markov process is weakly reversible, then the sum of the recurrence order and the cost are constants on each sets of states connected by recurrent arcs. This allows us to determine the necessary and sufficient conditions on the cooling rate for the optimization algorithm to hit a global minimum with probability one from all initial states.

REFERENCES

[1] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311-329.

[2] J. N. TSITSIKLIS, *Markov chains with rare transitions and simulated annealing*, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, August 1985; revised November 1985.

[3] S. GEMAN AND D. GEMAN, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Machine Intelligence, 6 (1984), pp. 721-741.

[4] D. MITRA, F. ROMEO, AND A. SANGIOVANNI-VINCENTELLI, *Convergence and finite-time behavior of simulated annealing*, Adv. in Appl. Probab., 18 (1986), pp. 747-771.

[5] B. GIDAS, *Non-stationary Markov chains and convergence of the annealing algorithm*, J. Statist. Phys., 39 (1985), pp. 73-131.

[6] K. L. CHUNG, *A Course in Probability Theory*, Academic Press, Orlando, FL, 1974.

[7] D. P. CONNORS AND P. R. KUMAR, *Balance of recurrence order in time-inhomogeneous Markov chains with application to simulated annealing*, Probab. Engrg. Inform. Sci., 2 (1988), pp. 157-184.

[8] D. P. CONNORS, *Balance of recurrence order in time-inhomogeneous Markov chains with application to simulated annealing*, Ph.D. thesis, University of Illinois, Urbana, IL, 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1462-1476, November 1989

013

STABILIZATION OF UNCERTAIN SYSTEMS WITH NORM BOUNDED UNCERTAINTY—A CONTROL LYAPUNOV FUNCTION APPROACH*

MARIO A. ROTEA† AND PRAMOD P. KHARGONEKAR‡

Abstract. A robust stabilization problem in a state-space setting is treated. It is assumed that the states are available for feedback. Using a fixed Lyapunov function approach (*quadratic stability*) it is shown that an open loop stabilizability condition is equivalent to the existence of a stabilizing memoryless linear state-feedback controller. As a consequence, it is shown that the existence of a quadratically stabilizing nonlinear time-varying dynamic state-feedback controller implies the existence of a quadratically stabilizing memoryless linear time-invariant state-feedback compensator.

Key words. robust stabilization, quadratic stability and stabilizability, control Lyapunov functions, H^∞ control theory

AMS(MOS) subject classification. 93C35

1. Introduction.

Consider the uncertain linear system

(Σ_u):

$$\frac{dx}{dt}(t) = A(\delta(t))x(t) + B(\delta(t))u(t) \quad \text{a.e. } t \in \mathbb{R} \tag{1.1}$$

where $\delta(t)$ is a vector of unknown parameters belonging to a compact set. In this paper we will be particularly interested in the special case of "norm-bounded time-varying uncertainty" given by

(Σ_{nu}):

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t) + Dw(t), \tag{1.2a}$$

$$e(t) = E_1 x(t) + E_2 u(t), \tag{1.2b}$$

$$w(t) = \Delta(t)e(t), \tag{1.2c}$$

$$\Delta(t) \in \mathbb{U} := \{U \in \mathbb{R}^{k \times p} : \|U\| \leqq 1\} \quad \text{a.e. } t \in \mathbb{R}. \tag{1.2d}$$

(For more precise and detailed descriptions, see §§ 2 and 3.) Uncertain linear systems of the form (1.2) have been investigated in [11], [15], [16], [18], and [23]. In this paper we will explore the problem of robust stabilization of the uncertain system (Σ_{nu}) using state feedback. There are two somewhat different but related approaches for studying this problem: (a) using the small gain theorem in combination with H^∞ control theory based synthesis procedures, and (b) using synthesis methods based on quadratic Lyapunov functions also known as quadratic stabilization theory. In both approaches, it is assumed that a linear time-invariant controller is to be designed. Under the assumption that a linear time-invariant controller is to be designed, it has been shown in [11] that these two apparently unrelated techniques are actually equivalent. It also has been shown how a stabilizing compensator can be obtained by solving an algebraic Riccati equation. Thus, under the assumption of linear time-invariant (dynamic) state feedback, the problem of quadratic stabilization has been completely resolved.

* Received by the editors August 22, 1988; accepted for publication (in revised form) January 5, 1989. This work was supported in part by National Science Foundation grant ECS-8451519, grants from Honeywell and GE, and by U.S. Air Force Office of Scientific Research grant AFOSR-88-0020.

† Department of Electrical Engineering and Center for Control Science and Dynamical Systems, University of Minnesota, Minneapolis, Minnesota 55455.

‡ Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122.

There is, however, no compelling reason to restrict our attention to linear time-invariant controllers. In fact, Petersen [14] gives an example of an uncertain system of the form (1.1) that can be quadratically stabilized by nonlinear state-feedback but for which there does not exist a stabilizing linear time-invariant state-feedback. On the other hand, Hollot and Barmish [5] show that if the input matrix B in the system (1.1) is fixed and known, then the system is quadratically stabilizable if and only if there exists a stabilizing linear time-invariant controller. Also, Petersen and Barmish [18] (for single input systems), and Petersen [16], show that if in (1.2) there is no uncertainty in the state matrix A (i.e., $E_1 = 0$), the system is quadratically stabilizable if and only if there exists a quadratically stabilizing linear time-invariant controller. In view of Petersen's example it is an interesting question whether for the uncertain system (Σ_{nu}) quadratic stabilizability implies the existence of a stabilizing linear time-invariant controller. This open question was also mentioned in [23].

In this paper we take an "open loop" approach to the problem of quadratic stabilization of the uncertain system (Σ_u) and, in particular, (Σ_{nu}). The concept of "control Lyapunov function" is introduced as an open loop definition of quadratic stabilizability. It is also a natural generalization, to the setting of uncertain systems, of an elementary Lyapunov function characterization of the well-known concept of stabilizability for finite-dimensional linear systems. Roughly speaking, for a positive definite matrix P, the function $v(x) = x'Px$ is called a control Lyapunov function if for each x in $\mathbb{R}^n$, there exists a control input u (that may be a function of δ) such that the derivative of $v(x)$ along solutions to (1.1) is strictly negative. In § 2 it is shown that if the system (Σ_u) is quadratically stabilizable using nonlinear time-varying controllers, then it admits a control Lyapunov function. Of course, this last fact also applies to the uncertain system (Σ_{nu}).

The main result of this paper shows that the uncertain system (1.2) *admits a control Lyapunov function if and only if there exists a quadratically stabilizing linear time-invariant controller of the form* $u = Kx$. A consequence of this approach using control Lyapunov functions is that if there exists a nonlinear time-varying quadratically stabilizing controller, then there exists a quadratically stabilizing controller of the form $u = Kx$.

These results fit nicely into a collection of recent results on the possible advantages of nonlinear time-varying controllers over linear time-invariant controllers for robust and H^∞ optimal control. See, for example, [4], [7], [9], [10], [12], [13], and [20]. In particular, our main results are consistent with the (qualitative) **Plant Uncertainty Principle** of Khargonekar and Poolla that in the present context states the following. *In robust control problems for linear time-invariant plants, nonlinear time-varying controllers yield no advantage over linear time-invariant controllers if the plant uncertainty is unstructured.* It will be intuitively obvious that the uncertainty considered in this paper is unstructured.

As mentioned above, if we restrict our attention to linear time-invariant controllers, then it is shown in [11] that the quadratic stabilization problem is mathematically equivalent to the standard problem in H^∞ control theory. It should be noted that this equivalence between quadratic stabilization and H^∞ optimization is not known to be true if we consider nonlinear time-varying controllers. Thus, our result on the equivalence between nonlinear time-varying controllers and linear time-invariant controllers cannot be obtained as a special case of earlier results of Khargonekar and Poolla [9] on such an equivalence for H^∞ optimal control.

The organization of this paper is as follows. In § 2 we discuss the concept of control Lyapunov functions. In § 3 we give the main results of the paper. It turns out that we need to prove a matrix-theoretic result (Theorem 3.15) to establish the main

results of this paper. Roughly speaking, if the "discriminant" of a quadratic form involving a self-adjoint (second-order) polynomial matrix is always negative, then for some choice of the independent variable the resulting matrix is sign definite. This result appears to be unknown in the matrix theory literature and generalizes a previous result of Petersen and Hollot (see, for example, [19]).

The notation used through this paper is fairly standard. $\mathbb{R}(X)$ and $\mathbb{N}(X)$ are used to denote the range and null spaces of a linear operator (matrix) X. The empty set is denoted by $\varnothing$, and if k is a positive integer, we use $\underline{k}$ to represent the set $\{1, 2, \cdots, k\}$. The identity matrix is denoted by I. For a constant real matrix X, X' denotes its transpose and $\|X\|$ the maximum singular value of X. Moreover, if X is symmetric, $\lambda_m(X)$ and $\lambda_M(X)$ denote its minimum and maximum eigenvalues, respectively.

2. Lyapunov functions and uncertain systems. Let Σ denote a continuous-time linear time-invariant system

(Σ):

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t), \qquad t \in \mathbb{R} \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ denotes the input vector, and A and B are real matrices of compatible dimensions. Let $P \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. We will say that the function $v(x) := x'Px$ is a *control Lyapunov function* for the linear system (Σ) if there exists $\alpha > 0$ such that for each $x \in \mathbb{R}^n$ there exists u (that may depend on x) $\in \mathbb{R}^m$ such that

$$L(x) := x'(A'P + PA)x + 2x'PBu \leqq -\alpha \|x\|^2.$$

Note that $L(x)$ is the derivative of $v(x)$ subject to (2.1).

LEMMA 2.2. *The system* (Σ) *admits a control Lyapunov function if and only if it is stabilizable.*

The proof of this result is left as an easy exercise for the reader. Thus, we may regard a control Lyapunov function as a Lyapunov stability type characterization of stabilizability. It can also be thought of as an open loop characterization of stabilizability, although such a characterization is not the most elementary. We would like to remark that this characterization of stabilizability is not new. In fact, Sontag in [22] has already used the notion of a control Lyapunov function as a characterization of asymptotic controllability for nonlinear systems. *The central aim of this paper is to generalize Lemma* 2.2 *to uncertain systems.*

The uncertain systems under consideration are described by state equations of the form

(Σ_u):

$$\frac{dx}{dt}(t) = A(\delta(t))x(t) + B(\delta(t))u(t) \quad \text{a.e. } t \in \mathbb{R} \tag{2.3}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ denotes the input vector, and the functions $A(.)$, $B(.)$ are assumed to be continuous real-valued matrix functions. The vector-valued function $\delta(.)$ represents the parameter uncertainty (possibly time-varying) and it is assumed to be a Lebesgue measurable function that satisfies $\delta(t) \in \mathbb{U}$ (compact set in $\mathbb{R}^k$), where $t \in \mathbb{R}$ almost everywhere. In the sequel such a function $\delta(.)$ will be called an admissible uncertainty.

As before, let $P \in \mathbb{R}^{n\times n}$ denote a positive definite matrix. We will say that the function $v(x) := x'Px$ is a *control Lyapunov function* for the uncertain system (Σ_u) if there exists $\alpha > 0$ such that for each pair $(x, \delta) \in \mathbb{R}^n \times \mathbb{U}$ there exists u (that may depend on x and $\delta) \in \mathbb{R}^m$ such that

$$L(x, \delta) := x'(PA(\delta) + A'(\delta)P)x + 2x'PB(\delta)u \leqq -\alpha \|x\|^2. \tag{2.4}$$

We regard the existence of a control Lyapunov function as an open loop notion of stabilizability for the uncertain system (Σ_u). In the remainder of this section we shall show that the existence of a control Lyapunov function is a necessary condition for the quadratic stabilizability (see, for example, [1] and [17]) of (Σ_u) by a nonlinear time-varying dynamic state feedback.

For the stabilization of the uncertain system (Σ_u), we will consider a nonlinear dynamic state-feedback controller (Σ_c) described by

(Σ_c):

$$\frac{dx_c}{dt}(t) = f(x(t), x_c(t), t), \tag{2.5a}$$

$$u(t) = h(x(t), x_c(t), t) \tag{2.5b}$$

where $t \in \mathbb{R}$, $x_c(t) \in \mathbb{R}^q$ denotes the state of the compensator, $h(x, x_c, t) \in \mathbb{R}^m$, $f(x, x_c, t) \in \mathbb{R}^q$, and the functions $f(.)$ and $h(.)$ are such that (2.5) is a well-defined dynamical system. Note that the closed loop system (Σ_{cl}) obtained from the feedback interconnection of the uncertain system (Σ_u) and the compensator (Σ_c) can be described by the following equations:

(Σ_{cl}):

$$\frac{dz}{dt}(t) = F(t)z(t) + G(t)p(z(t), t) \quad \text{a.e. } t \in \mathbb{R} \tag{2.6a}$$

where $z(t) \in \mathbb{R}^{n+q}$ denotes the composite state $[x'(t)\ x_c'(t)]'$, and

$$F(t) := \begin{bmatrix} A(\delta(t)) & 0 \\ 0 & 0 \end{bmatrix}, \tag{2.6b}$$

$$G(t) := \begin{bmatrix} B(\delta(t)) & 0 \\ 0 & I \end{bmatrix}, \tag{2.6c}$$

$$p(z, t) := \begin{bmatrix} h(x(t), x_c(t), t) \\ f(x(t), x_c(t), t) \end{bmatrix}. \tag{2.6d}$$

The stability of the closed loop system (Σ_{cl}) will be studied using Lyapunov stability theory. We will consider the case where the Lyapunov function is quadratic. More precisely, we have the following definition (see also [1] and [17]).

The uncertain system (Σ_u) is said to be *quadratically stabilizable* if:

(i) There exists an integer $q \geqq 0$, a feedback control law $p(.): \mathbb{R}^{n+q} \times \mathbb{R} \to \mathbb{R}^{m+q}$, with $p(0, t) = 0$ for all $t \in \mathbb{R}$, such that $p(., t)$ is continuous and $p(x, x_c, .)$ is an (essentially) bounded measurable function over $\mathbb{R}$; and

(ii) A symmetric positive definite matrix $P \in \mathbb{R}^{(n+q)\times(n+q)}$, and a constant $\alpha > 0$, such that the following condition is satisfied. Given any admissible uncertainty $\delta(.)$, the Lyapunov derivative, corresponding to the closed loop system (Σ_{cl}) and the Lyapunov function $z'Pz$, satisfies the inequality

$$L(z, t) := z'(PF(t) + F'(t)P)z + 2z'PG(t)p(z, t) \leqq -\alpha \|z\|^2 \quad \text{a.e. } t \in \mathbb{R}, \tag{2.7}$$

and for all $z \in \mathbb{R}^{n+q}$.

Finally, suppose that the system (Σ_u) is quadratically stabilizable and, in addition, the stabilizing control law can be chosen to be $u(t) = Kx(t)$, for some real matrix K; then, (Σ_u) is said to be *quadratically stabilizable via (memoryless) time-invariant linear control.*

Note that if the uncertain system (Σ_u) is quadratically stabilizable, then $z'Pz$ is a Lyapunov function for the closed loop system (2.6). As is well known, given any $t_0 \in \mathbb{R}$, any initial condition $z(t_0) = z_0$, and an admissible uncertainty $\delta(.)$, there exists $t_1 > t_0$ such that (2.6a) admits a unique solution on $[t_0, t_1]$, and such a solution is continuable over $[t_0, \infty)$. Moreover, in the view of (2.7), the equilibrium point $z_\infty = 0$ will be uniformly asymptotically stable in the large (see, for example, [1] and [17]).

Now we are ready to show the connection between the control Lyapunov function concept and the notion of quadratic stabilizability.

PROPOSITION 2.8. *Consider the uncertain system* (Σ_u) *defined by* (2.3). *If* (Σ_u) *is quadratically stabilizable, then* (Σ_u) *admits a control Lyapunov function.*

Proof. Suppose that the system (Σ_u) is quadratically stabilizable. Then, there exists an integer $q \geqq 0$, a control law $p(.): \mathbb{R}^{n+q} \times \mathbb{R} \to \mathbb{R}^{m+q}$, a symmetric positive definite matrix $P \in \mathbb{R}^{(n+q)\times(n+q)}$, and a constant $\alpha > 0$, such that (2.7) holds for any admissible uncertainty $\delta(.)$. If $q = 0$, the conclusion is obvious. Indeed, $v(x) = x'Px$ is a control Lyapunov function for the system (Σ_u). If q is nonzero, we proceed as follows.

Define the symmetric positive definite matrix $S := P^{-1}$ and let the change of coordinates $z = Sw$ be applied to (2.7). Then, it follows that

$$w'(F(t)S + SF'(t))w + 2w'G(t)\tilde{p}(w, t) \leqq -\tilde{\alpha}\|w\|^2 \quad \text{a.e. } t \in \mathbb{R}, \tag{2.9}$$

for all $w \in \mathbb{R}^{n+q}$, and some $\tilde{\alpha} > 0$, where $\tilde{p}(w, t) = p(z, t)$. In particular, for $w = [r' \ 0]'$, $r \in \mathbb{R}^n$, from (2.6), (2.9), and partitioning S as follows:

$$S = \begin{bmatrix} S_1 & S_2 \\ S_1' & S_3 \end{bmatrix}$$

where the dimension of S_1 is $n \times n$, the dimension of S_2 is $n \times q$, and the dimension of S_3 is $q \times q$, we obtain

$$r'(A(\delta(t))S_1 + S_1A'(\delta(t)))r + 2r'B(\delta(t))\tilde{h}(r, t) \leqq -\tilde{\alpha}\|r\|^2 \quad \text{a.e. } t \in \mathbb{R}, \tag{2.10}$$

for all $r \in \mathbb{R}^n$, and any admissible uncertainty $\delta(.)$, where $\tilde{h}(r, t) = h(S_1r, S_2'r, t)$.

Finally, considering (2.10) for a time-invariant uncertainty $\delta \in \mathbb{U}$, setting $t = t^*$ (any fixed time such that $\tilde{h}(., t^*)$ is well defined), and with the coordinate transformation $x = S_1r$, we conclude that there exists $\beta > 0$ such that

$$L(x, \delta) := x'(P_1A(\delta) + A'(\delta)P_1)x + 2x'P_1B(\delta)h^*(x) \leqq -\beta\|x\|^2,$$

for all $x \in \mathbb{R}^n$ and $\delta \in \mathbb{U}$, where P_1 is a symmetric positive definite matrix given by $P_1 := S_1^{-1}$ and $h^*(x) = \tilde{h}(P_1x, t^*)$. The proof is concluded observing that this last inequality implies that the function $v(x) := x'P_1x$ is a control Lyapunov function for the uncertain system (Σ_u). □

In [17], Petersen has obtained a result similar to Proposition 2.8, which states that if an uncertain system (Σ_u) (as defined in (2.3)) is quadratically stabilizable, then it is quadratically stabilizable via memoryless *nonlinear* state feedback. Note that, in general, the notion of control Lyapunov function is not equivalent to the notion of quadratic stabilization via memoryless compensators. In fact, the former is much weaker since the control input "u" is allowed to be a function of the uncertainty "δ" (see, for example, (2.4)).

3. Control Lyapunov functions and stabilizability by linear feedback. In this section the case of norm-bounded time-varying uncertainty is considered. Such an uncertainty representation has already been used by Petersen [15] and [16], Zhou and Khargonekar [23], and Khargonekar, Petersen, and Zhou [11], to develop necessary and sufficient conditions for the existence of linear (quadratically) stabilizing compensators not only for the state feedback case but also for output feedback. Here, we will show that those conditions are actually necessary for the existence of a control Lyapunov function. As a consequence, from Proposition 2.8 we conclude that, for the case of norm-bounded time-varying uncertainty, nonlinear time-varying dynamic compensators are not better than linear (memoryless) ones when the states are available for feedback and a single quadratic Lyapunov function is sought to establish the stability of the closed loop system.

A system (Σ_{nu}) with norm-bounded uncertainty is better described by the model given below rather than the general representation given in (2.3):

(Σ_{nu}):

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t) + Dw(t), \tag{3.1a}$$

$$e(t) = E_1 x(t) + E_2 u(t), \tag{3.1b}$$

$$w(t) = \Delta(t) e(t), \tag{3.1c}$$

where $t \in \mathbb{R}$ almost everywhere, $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^m$ denotes the input vector, $w(t) \in \mathbb{R}^k$, and $e(t) \in \mathbb{R}^p$. The real matrices A, B, D, E_1, and E_2 are known and of appropriate dimensions. As before, the matrix-valued function $\Delta(.)$ is assumed to be Lebesgue measurable and satisfies

$$\Delta(t) \in \mathbb{U} := \{U \in \mathbb{R}^{k \times p} : \|U\| \leqq 1\} \quad \text{a.e. } t \in \mathbb{R}. \tag{3.1d}$$

Furthermore, to avoid trivial situations, it will be assumed that the dimension of the input is strictly less than the dimension of the state (i.e., $m < n$).

The main result of this paper is the following theorem.

THEOREM 3.2. *Consider the uncertain system* (Σ_{nu}) *defined in* (3.1). *Then, the following statements are equivalent*:

(i) *The uncertain system* (Σ_{nu}) *admits a control Lyapunov function.*

(ii) *The uncertain system* (Σ_{nu}) *is quadratically stabilizable.*

(iii) *The uncertain system* (Σ_{nu}) *is quadratically stabilizable via linear time-invariant* (*memoryless*) *control.*

It is obvious that (iii) implies (ii). In Proposition 2.8 it has been proved (in a more general situation) that (ii) implies (i). We will prove that (i) $\Rightarrow$ (iii). The proof is rather long and difficult. For ease of exposition, some intermediate results will be developed, leading to a proof of Theorem 3.2.

First note that if the uncertain system (Σ_{nu}) admits a control Lyapunov function, it readily follows from (2.4) and (3.1) that there exists a symmetric positive definite matrix $P \in \mathbb{R}^{n \times n}$ such that the following condition holds. Given any nonzero $x \in \mathbb{R}^n$, if

$$(B + D\Delta E_2)' Px = 0 \quad \text{for some } \Delta \in \mathbb{U}, \tag{3.3a}$$

then

$$x'(A'P + PA)x + 2x'(PD\Delta E_1)x < 0. \tag{3.3b}$$

Let $z \in \mathbb{R}^n$ be given and define the set

$$\Gamma(z) := \{\Gamma \in \mathbb{R}^{p \times k} : (B + D\Gamma' E_2)' z = 0, \|\Gamma\| \leqq 1\}, \tag{3.4}$$

and note that, when $\Gamma(z) \neq \emptyset$, it is a compact set. Then, from the discussion above we conclude that given any nonzero $z \in \mathbb{R}^n$, if $\Gamma(z) \neq \emptyset$, the following inequality holds:

$$\max \{z'(AS+SA')z + 2z'(D\Gamma' E_1 S)z : \Gamma \in \Gamma(z)\} < 0 \tag{3.5}$$

where $S := P^{-1}$ is a symmetric positive definite matrix. Indeed, making the change of coordinates $z = Px$ in (3.3) it follows that Δ satisfies (3.3a) if and only if $\Gamma = \Delta' \in \Gamma(z)$. Moreover, the left-hand side of (3.3b) is continuous in $\Gamma = \Delta'$.

It should be noted that the situation $\Gamma(z) = \emptyset$ for all nonzero $z \in \mathbb{R}^n$ is rather uninteresting. Indeed, the latter will never occur when the dimension of the input space (m) is less than the dimension of the state space (n). In the next lemma the set $\Gamma(z)$ is studied and the maximization problem introduced in (3.5) is solved.

For a matrix $W \in \mathbb{R}^{m \times p}$ the matrix $W^+ \in \mathbb{R}^{p \times m}$ will denote the Moore–Penrose inverse of W (see, for example [21, § 3.3]). Note that W^+ is the unique matrix that satisfies

$$WW^+W = W, \qquad (W^+W)' = W^+W, \tag{3.6a}$$

$$W^+WW^+ = W^+, \qquad (WW^+)' = WW^+. \tag{3.6b}$$

Conditions (3.6) amount to the requirement that

$$\Pi_1 := W^+W \quad \text{and} \quad \Pi_2 := WW^+, \tag{3.6c}$$

be the orthogonal projections onto $\mathbb{R}(W')$ and $\mathbb{R}(W)$, respectively.

LEMMA 3.7. *Let $W \in \mathbb{R}^{m \times p}$, $W \neq 0$; $d \in \mathbb{R}^k$; $b \in \mathbb{R}^m$; and $e \in \mathbb{R}^p$ be given matrices. Define the following set:*

$$\Gamma := \{\Gamma \in \mathbb{R}^{p \times k} : W\Gamma d = b, \|\Gamma\| \leqq 1\}. \tag{3.8}$$

Then, $\Gamma \neq \emptyset$ if and only if $b \in \mathbb{R}(W)$ and $\|W^+b\| \leqq \|d\|$, where $W^+ \in \mathbb{R}^{p \times m}$ is the Moore–Penrose inverse of W. Moreover, if $\Gamma \neq \emptyset$ and Π_1 is the matrix defined in (3.6c), *then*

$$\max \{e'\Gamma d : \Gamma \in \Gamma\} = e'W^+b + \|(I - \Pi_1)e\| \{\|d\|^2 - \|W^+b\|^2\}^{1/2}. \tag{3.9}$$

Proof. The proof is in two steps. First, we establish the necessary and sufficient conditions for the nonemptiness of the set Γ.

(Necessity.) Suppose that there exists Γ such that $W\Gamma d = b$ and $\|\Gamma\| \leqq 1$. Obviously, $b \in \mathbb{R}(W)$, and, from (3.6), $W^+b = \Pi_1 \Gamma d$. Since Π_1 is an orthogonal projection and $\|\Gamma\| \leqq 1$, the latter implies $\|W^+b\| \leqq \|\Pi_1\| \|\Gamma\| \|d\| \leqq \|d\|$.

(Sufficiency.) Suppose that $b \in \mathbb{R}(W)$ and $\|W^+b\| \leqq \|d\|$. We now consider two cases.

Case 1. $d = 0$. Obviously $\Gamma \neq \emptyset$ if and only if $b = 0$. Now, from the assumptions it follows that $b \in \mathbb{R}(W)$ and $b \in \mathbb{N}(W^+)$. When (3.6) is used, it follows that $\mathbb{N}(W^+) = \{\mathbb{R}(W)\}^\perp$. Hence, we conclude that $b = 0$.

Case 2. $d \neq 0$. Consider the following matrices:

$$\Gamma_0 := W^+bd^+, \qquad d^+ := (d'd)^{-1}d'. \tag{3.10}$$

It will be shown that $\Gamma_0 \in \Gamma$. From (3.6) and since $b \in \mathbb{R}(W)$ we obtain

$$\begin{aligned} W\Gamma_0 d &= WW^+b = WW^+Wc, \qquad c \in \mathbb{R}^p \\ &= Wc = b. \end{aligned}$$

Moreover, $\|\Gamma_0\| \leqq \|W^+b\| \|d^+\| \leqq \|W^+b\| / \|d\| \leqq 1$. This concludes the first part of the proof. We are now ready to establish (3.9).

Note first that, when $\mathbb{\Gamma} \neq \varnothing$, the right-hand side of (3.9) is well defined. We now consider two cases.

Case 1. $d = 0$. In this case $\mathbb{\Gamma} \neq \varnothing$ implies $b = 0$. The result follows by inspection.

Case 2. $d \neq 0$. In this case $\mathbb{\Gamma} \neq \varnothing$ implies $b \in \mathbb{R}(W)$ and $\|W^+ b\| \leqq \|d\|$. Let $e \in \mathbb{R}^p$ denote an arbitrary vector. First we show that the inequality

$$e'\Gamma d \leqq e' W^+ b + \|(I - \Pi_1)e\| \{\|d\|^2 - \|W^+ b\|^2\}^{1/2}, \tag{3.11}$$

holds for any $\Gamma \in \mathbb{\Gamma}$.

Let $\Gamma \in \mathbb{\Gamma}$ be given. Then Γ satisfies the equation $W\Gamma d = b$. Therefore, there exists $Z \in \mathbb{R}^{p \times k}$ such that (see, for example, [21, § 2.3])

$$\Gamma = \Gamma_0 + Z - \Pi_1 Z d d^+$$

where Γ_0 and d^+ were defined in (3.10). Thus, we obtain

$$\Gamma d = W^+ b + (I - \Pi_1) Z d. \tag{3.12}$$

Moreover, for any $Z \in \mathbb{R}^{p \times k}$, it follows from (3.6) that

$$(W^+ b)'(I - \Pi_1) Z d = 0. \tag{3.13}$$

Hence, (3.12), (3.13), and the fact that $\|\Gamma\| \leqq 1$, imply the following chain of inequalities:

$$\begin{aligned} e'\Gamma d &= e' W^+ b + e'(I - \Pi_1) Z d \\ &= e' W^+ b + ((I - \Pi_1)e)'(I - \Pi_1) Z d \\ &\leqq e' W^+ b + \|(I - \Pi_1)e\| \, \|(I - \Pi_1) Z d\| \\ &\leqq e' W^+ b + \|(I - \Pi_1)e\| \{\|\Gamma d\|^2 - \|W^+ b\|^2\}^{1/2} \\ &\leqq e' W^+ b + \|(I - \Pi_1)e\| \{\|d\|^2 - \|W^+ b\|^2\}^{1/2}. \end{aligned}$$

Since $\Gamma \in \mathbb{\Gamma}$ is arbitrary, this last row implies inequality (3.11).

To complete the proof we must show that the upper bound given in (3.11) is actually achieved. To that purpose we define

$$\Gamma^* := \begin{cases} \Gamma_0 + \gamma (I - \Pi_1) e d^+ & \text{if } (I - \Pi_1)e \neq 0, \\ \Gamma_0 & \text{otherwise} \end{cases} \tag{3.14a}$$

where Γ_0 and d^+ are as in (3.10), and

$$\gamma := \frac{\{\|d\|^2 - \|W^+ b\|^2\}^{1/2}}{\|(I - \Pi_1)e\|}. \tag{3.14b}$$

Note first that $\Gamma^* \in \mathbb{\Gamma}$. The case $\Gamma^* = \Gamma_0$ already has been done in the first part of the proof. When $(I - \Pi_1)e \neq 0$, a trivial computation using (3.6) and (3.14a) will show that $W\Gamma^* d = b$. To show that Γ^* is a contraction, we proceed as follows. Let $x \in \mathbb{R}^k$ be given. From (3.14a) we obtain

$$\begin{aligned} \Gamma^* x &= W^+ b d^+ x + \gamma (I - \Pi_1) e d^+ x \\ &= [W^+ b + \gamma (I - \Pi_1) e] d^+ x. \end{aligned}$$

When we use (3.13) and (3.14b), it follows that

$$\|\Gamma^* x\|^2 = \|d\|^2 (d^+ x)^2 \leqq \|d\|^2 \|d^+\|^2 \|x\|^2 = \|x\|^2.$$

Since x is arbitrary we conclude that $\|\Gamma^*\| \leqq 1$.

Finally, an easy calculation shows that $e'\Gamma^* d$ achieves the upper bound (3.11). □

The next result generalizes a previous result of Petersen and Hollot [19], and it will be used in the proof of Theorem 3.2. For the sake of clarity, its proof is given in a separate Appendix.

THEOREM 3.15. *Let* $A_1 = A_1'$, $A_2 = A_2' \geqq 0$, *and* $A_3 = A_3'$ *denote* $s \times s$ *real matrices. Suppose that for all* $x \neq 0$ *such that* $x'A_3x \geqq 0$, *we have*

(i) $x'A_1x < 0$; *and*

(ii) $\delta(x) := (x'A_1x)^2 - 4(x'A_2x)(x'A_3x) > 0$.

Then, there exists $\beta > 0$ *such that* $\beta^2 A_2 + \beta A_1 + A_3 < 0$.

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. Consider the uncertain system (Σ_{nu}) defined in (3.1) and suppose that it admits a control Lyapunov function. It will be shown that (Σ_{nu}) is quadratically stabilizable via (memoryless) linear time-invariant control. Let us assume first that $E_2 \neq 0$. From the discussion below Theorem 3.2, it follows that there exists a symmetric positive definite matrix $S \in \mathbb{R}^{n\times n}$ such that given any nonzero $z \in \mathbb{R}^n$, if $\mathbb{\Gamma}(z) \neq \varnothing$ (see (3.4)), then inequality (3.5) holds.

With reference to the notation of Lemma 3.7, we now make the following association:

$$E_2' \leftrightarrow W,$$

$$D'z \leftrightarrow d,$$

$$-B'z \leftrightarrow b,$$

$$E_1Sz \leftrightarrow e,$$

$$\mathbb{\Gamma}(z) \leftrightarrow \Gamma,$$

and (as before) let $W^+ \in \mathbb{R}^{p\times m}$ denote the Moore–Penrose inverse of W and Π_1, Π_2 be defined by (3.6c). It now follows from (3.5) and Lemma 3.7 that for each $z \neq 0$, such that

$$B'z \in \mathbb{R}(W) \quad \text{and} \quad \|W^+B'z\| \leqq \|D'z\|, \tag{3.16a}$$

the following inequality holds:

$$\begin{aligned} z'(AS + SA')z - 2(E_1Sz)'W^+B'z \\ + 2\|(I - \Pi_1)E_1Sz\|\{\|D'z\|^2 - \|W^+B'z\|^2\}^{1/2} < 0. \end{aligned} \tag{3.16b}$$

Let $A_0 \in \mathbb{R}^{n\times n}$ be defined by

$$A_0 := A - BW^{+\prime}E_1, \tag{3.17}$$

and observe that from (3.6c) we obtain that $\mathbb{N}(I - \Pi_2) = \mathbb{R}(W)$. Hence, we may conclude that (3.16b) holds for any $z \neq 0$ such that $\|W^+B'z\| \leqq \|D'z\|$ and $z \in \mathbb{N}((I - \Pi_2)B')$.

Let X denote a (full column rank) real matrix whose columns span $\mathbb{N}((I - \Pi_2)B')$, and define the following symmetric matrices:

$$A_1 := X'(A_0S + SA_0')X, \quad A_0 \text{ defined in (3.17)},$$

$$A_2 := X'SE_1'(I - \Pi_1)E_1SX,$$

$$A_3 := X'(DD' - BW^{+\prime}W^+B')X.$$

Note that $A_2 \geqq 0$ (for $I - \Pi_1$ is an orthogonal projection). It follows from (3.16) that

$$y'A_1y + 2\{(y'A_2y)(y'A_3y)\}^{1/2} < 0,$$

for all $y \neq 0$ such that $y'A_3y \geqq 0$. Or equivalently, $y'A_1y < 0$ and $(y'A_1y)^2 - 4(y'A_2y) \times (y'A_3y) > 0$, for all $y \neq 0$ such that $y'A_3y \geqq 0$. Using Theorem 3.15 we may conclude that there exists $\beta > 0$ such that $\beta^2 A_2 + \beta A_1 + A_3 < 0$. Hence, letting $S_0 := \beta S > 0$ and $E_0 := (I - \Pi_1)E_1$, it follows that

$$z'(S_0E_0'E_0S_0 + A_0S_0 + S_0A_0' + DD' - BW^{+\prime}W^+B')z < 0, \tag{3.18}$$

for any $z \neq 0$ such that $z \in \mathbb{N}((I - \Pi_2)B')$.

Now, from (3.18), a standard argument using the Finsler Lemma (see, for example, [6, § 3.2.6]) shows that there exists $\varepsilon > 0$ such that the matrix

$$Q_0 := -\left(S_0E_0'E_0S_0 + A_0S_0 + S_0A_0' + DD' - BW^{+\prime}W^+B' - \frac{1}{\varepsilon} B(I - \Pi_2)B' \right),$$

is positive definite. Letting $P_0 := S_0^{-1} > 0$ and $Q := P_0Q_0P_0 > 0$, we conclude that the following Riccati equation holds:

$$A_0'P_0 + P_0A_0 + E_0'E_0 + P_0\left(DD' - BW^{+\prime}W^+B' - \frac{1}{\varepsilon} B(I - \Pi_2)B' \right)P_0 + Q = 0. \tag{3.19}$$

Consider the linear time-invariant memoryless feedback law given by

$$Y := \frac{1}{2\varepsilon}(I - \Pi_2) + W^{+\prime}W^+, \tag{3.20a}$$

$$u = p(x) := -YB'P_0x - W^{+\prime}E_1x. \tag{3.20b}$$

We will next show that $v(x) := x'P_0x$ is a Lyapunov function for the closed loop system resulting from the feedback interconnection of the uncertain system (Σ_{nu}) and the controller (3.20). In fact, given any admissible uncertainty $\Delta(.)$, the derivative of $v(x)$ along the system (Σ_{nu}) defined in (3.1) subject to the control law (3.20) is given by

$$L(x, t) := x'(P_0A + A'P_0)x + 2x'P_0Bp(x) + 2x'P_0D\Delta(t)(E_1x + E_2p(x)) \tag{3.21}$$

where $\Delta(t) \in \mathbb{U}$. Hence, from (3.21) we obtain

$$\begin{aligned} L(x, t) &\leqq x'(P_0A + A'P_0)x + 2x'P_0Bp(x) + 2\|D'P_0x\| \, \|E_1x + E_2p(x)\|, \\ &\leqq x'(P_0A + A'P_0)x + 2x'P_0Bp(x) + \|D'P_0x\|^2 \\ &\quad + \|E_1x + E_2p(x)\|^2 \quad \text{a.e. } t \in \mathbb{R}. \end{aligned} \tag{3.22}$$

Substituting (3.17) and (3.20) in (3.22), and using (3.6), it follows that

$$\begin{aligned} L(x, t) \leqq x'\Big(& P_0A_0 + A_0'P_0 + E_0'E_0 + P_0DD'P_0 - P_0BW^{+\prime}W^+B'P_0 - \frac{1}{\varepsilon} P_0B(I - \Pi_2)B'P_0 \Big)x \\ & - \|W^+B'P_0x\|^2 + \|W'YB'P_0x\|^2. \end{aligned}$$

From (3.19) and this last inequality we may conclude that

$$L(x, t) \leqq -x'Qx - \|W^+B'P_0x\|^2 + \|W'YB'P_0x\|^2. \tag{3.23}$$

Finally, substituting (3.20a) in (3.23) and since $W'Y = \Pi_1W^+$, we obtain that

$$\begin{aligned} L(x, t) &\leqq -x'Qx - \|W^+B'P_0x\|^2 + \|\Pi_1W^+B'P_0x\|^2 \\ &\leqq -x'Qx + (\|\Pi_1\|^2 - 1)\|W^+B'P_0x\|^2 \leqq -\lambda_m(Q)\|x\|^2, \end{aligned}$$

where $t \in \mathbb{R}$ almost everywhere, and for all $x \in \mathbb{R}^n$. Since $\lambda_m(Q) > 0$, the proof for the case $E_2 \neq 0$ has been completed.

Finally, suppose that $E_2 = 0$. From (3.3) it follows that there exists a positive definite matrix $P \in \mathbb{R}^{n\times n}$ such that

$$x'(A'P + PA)x + 2x'(PD\Delta E_1)x < 0 \tag{3.24}$$

for all nonzero $x \in \mathbb{N}(B'P)$ and for all $\Delta \in \mathbb{U}$. Maximizing (3.24) over all possible $\Delta \in \mathbb{U}$ as in [15] (see also [2]), we may conclude that there exist $P_0 > 0$ and $\varepsilon > 0$ such that the Riccati equation (3.19) holds, provided that we adopt $W^+ = 0$. The rest of the proof remains the same as the case $E_2 \neq 0$. □

The next result gives a necessary and sufficient open loop condition to test whether the uncertain system (3.1) is quadratically stabilizable. A restricted version of this result was first given by Khargonekar, Petersen, and Zhou in [11], within the context of linear time-invariant compensators. Consider the uncertain system (Σ_{nu}) defined in (3.1) and set $W := E_2'$. Let W^+, Π_1, and Π_2 be defined by (3.6) and define $A_0 := A - BW^{+\prime}E_1$. If $E_2 = 0$, we set $W^+ = 0$.

COROLLARY 3.25. *The uncertain system* (Σ_{nu}) *is quadratically stabilizable if there exists* $\varepsilon > 0$ *such that the Riccati equation*

$$A_0'P + PA_0 + P\left(DD' - BW^{+\prime}W^+B' - \frac{1}{\varepsilon}B(I-\Pi_2)B'\right)P + E_1'(I-\Pi_1)E_1 + \varepsilon I = 0 \tag{3.26}$$

has a symmetric positive definite solution $P \in \mathbb{R}^{n\times n}$. *Furthermore, if such a solution exists,*

$$u := -\left[\frac{1}{2\varepsilon}(I-\Pi_2) + W^{+\prime}W^+\right]B'Px - W^{+\prime}E_1x, \tag{3.27}$$

is a quadratically stabilizing linear time-invariant control law. Conversely, if the uncertain system (Σ_{nu}) *is quadratically stabilizable, then there exists* $\varepsilon_1 > 0$ *such that for all* ε *in* $(0, \varepsilon_1]$, *the Riccati equation* (3.26) *has a unique positive definite solution* P_0 *such that* $[A_0 + (DD' - BW^{+\prime}W^+B' - (1/\varepsilon)B(I-\Pi_2)B')P_0]$ *is asymptotically stable.*

(A solution P_0 such that $[A_0 + (DD' - BW^{+\prime}W^+B' - (1/\varepsilon)B(I-\Pi_2)B')P_0]$ is asymptotically stable is called the *stabilizing solution.*)

Corollary 3.25 follows immediately from Theorem 3.2 and earlier results of Khargonekar, Petersen, and Zhou in [11]. This corollary also leads to the following conceptual algorithm for checking quadratic stabilizability.

(i) Set ε to some starting value; i.e., $\varepsilon = 1$.

(ii) Find the unique stabilizing solution to the Riccati equation (3.26) using any standard algorithm. If this solution exists and it is positive definite, stop; the system is quadratically stabilizable (and a stabilizing compensator is given by (3.27)). Else go to step (iii) below.

(iii) Replace ε by $\varepsilon/2$. If ε is less than the computational accuracy, stop; the system is not quadratically stabilizable. Else repeat step (ii) above.

Finally, it should be noted that to test whether the uncertain system (Σ_{nu}) is quadratically stabilizable, a one-parameter search must be performed. This is not the case if $E_2'E_2$ is nonsingular. In fact, in this case $W^+ = E_2(E_2'E_2)^{-1}$ and $\Pi_2 = I$. Furthermore, from the connection established in [11] between quadratic stability and the Small Gain Theorem, and using the recent results of Doyle et al. [3], it follows that, when $E_2'E_2$ is nonsingular, Corollary 3.25 can be strengthened to the following result.

COROLLARY 3.28. *Consider the uncertain system* (Σ_{nu}) *defined in* (3.1). *Suppose* $\Theta := E_2'E_2$ *is nonsingular,* $\{(I - E_2\Theta^{-1}E_2')E_1, A - B\Theta^{-1}E_2'E_1\}$ *is observable, and* $\{A, B\}$

is stabilizable. Then, the uncertain system is quadratically stabilizable if and only if the Riccati equation

$$(A-B\Theta^{-1}E_2'E_1)'P+P(A-B\Theta^{-1}E_2'E_1)+P(DD'-B\Theta^{-1}B')P + E_1'(I-E_2\Theta^{-1}E_2')E_1=0 \tag{3.29}$$

admits a (unique) symmetric positive definite solution $P_0\in\mathbb{R}^{n\times n}$ *such that* $[A-B\Theta^{-1}E_2'E_1+(DD'-B\Theta^{-1}B')P_0]$ *is asymptotically stable. Furthermore, if such a solution exists,*

$$u:=-\Theta^{-1}[B'P_0+E_2'E_1]x, \tag{3.30}$$

is a quadratically stabilizing linear time-invariant control law.

Finally, the existence of the unique stabilizing solution P_0 in the above results can be checked and, if it exists, P_0 can be computed using the associated Hamiltonian matrix.

Appendix. First we establish the following weaker version of Theorem 3.15.

LEMMA A1. *Let* $A_1=A_1'$, $A_2=A_2'>0$, *and* $A_3=A_3'$ *denote* $s\times s$ *real matrices. Suppose the following*:

(i) *For all* $x\neq 0$, $\delta(x):=(x'A_1x)^2-4(x'A_2x)(x'A_3x)>0$; *and*

(ii) *For all* $x\neq 0$, *such that* $x'A_3x\geqq 0$, $x'A_1x<0$.

Then, there exists $\beta>0$ *such that*

$$M(\beta):=\beta^2A_2+\beta A_1+A_3\leqq 0. \tag{A2}$$

Proof. Without loss of generality we can assume $A_2=I$. For, if this is not the case, we can make the change of coordinates $z=A_2^{1/2}x$ to reduce the problem to one with $A_2=I$. (Note that the symmetry of A_1 and A_3 as well as assumptions (i) and (ii) above remain invariant under this coordinate transformation.)

Suppose that (A2) does not hold. Then, for each $\beta>0$ there exists $x(\beta)\neq 0$ such that

$$\phi(\beta):=\beta^2(x'(\beta)x(\beta))+\beta(x'(\beta)A_1x(\beta))+(x'(\beta)A_3x(\beta))>0. \tag{A3}$$

It will be shown later that $x(\beta)$ can be chosen to be a continuous function of β on the interval (β_0,∞), for some $\beta_0<0$ such that $\|x(\beta)\|=1$ and (A3) holds for all $\beta>0$.

Since for each $x\neq 0$, $\delta(x)>0$ it follows that

$$\rho_1(\beta):=\tfrac{1}{2}(-x'(\beta)A_1x(\beta)-\sqrt{\delta(x(\beta))}), \tag{A4a}$$

$$\rho_2(\beta):=\tfrac{1}{2}(-x'(\beta)A_1x(\beta)+\sqrt{\delta(x(\beta))}), \tag{A4b}$$

are real-valued functions and $\rho_2(\beta)>\rho_1(\beta)$ for every $\beta>0$. Moreover, the continuity of the mapping $\beta\to x(\beta)$ implies the continuity of the mappings $\beta\to\rho_1(\beta)$ and $\beta\to\rho_2(\beta)$ on the interval (β_0,∞). From (A3) and (A4) we must have

$$(\beta-\rho_1(\beta))(\beta-\rho_2(\beta))>0 \quad \text{for all } \beta>0. \tag{A5}$$

From (A5), it follows that given any $\beta^*>0$, either $\beta^*<\rho_1(\beta^*)<\rho_2(\beta^*)$ or $\beta^*>\rho_2(\beta^*)>\rho_1(\beta^*)$. Suppose first that for some $\beta^*>0$ we have $\beta^*<\rho_1(\beta^*)<\rho_2(\beta^*)$. Since $\beta-\rho_1(\beta)$ is continuous on the interval (β_0,∞), we observe that this function cannot change its sign over $(0,\infty)$. For if this is not the case, (A5) would be contradicted. Hence, it is concluded that $\beta-\rho_1(\beta)<0$, for all $\beta>0\Rightarrow 2\beta+x'(\beta)A_1x(\beta)<0$, for all $\beta>0\Rightarrow 2\beta+\lambda_m(A_1)<0$, for all $\beta>0$, which leads to the obvious contradiction $2\leqq 0$.

Now suppose that for some $\beta^* > 0$ we have $\beta^* > \rho_2(\beta^*) > \rho_1(\beta^*)$. Arguing as before, we conclude that

$$\beta - \rho_2(\beta) > 0, \text{ for all } \beta > 0 \Rightarrow 2\beta + x'(\beta)A_1 x(\beta) > \sqrt{\delta(x(\beta))}, \text{ for all } \beta > 0$$
$$\Rightarrow x_0' A_1 x_0 \geqq \sqrt{\delta(x_0)}$$

where $x_0 = \lim_{\beta \to 0} x(\beta) \neq 0$. (Note that this limit exists for $\beta \to x(\beta)$ continuous on (β_0, ∞), for some β_0 negative.) From the last inequality it follows that for such a nonzero x_0, $x_0' A_1 x_0 > 0$ and $x_0' A_3 x_0 \geqq 0$, which contradicts assumption (ii).

To complete the proof we must show that $x(\beta)$ can be chosen to be a continuous function of β. For each $\beta \in \mathbb{R}$, the matrix $M(\beta)$ defined in (A2) is symmetric, and therefore has real eigenvalues and s real orthonormal eigenvectors. That is, for each $\beta \in \mathbb{R}$ we can write

(A6a) $$M(\beta) = U(\beta)\Lambda(\beta)U'(\beta),$$

(A6b) $$\Lambda(\beta) = \operatorname{diag}(\lambda_j(\beta); j \in \underline{s}),$$

(A6c) $$U(\beta) = [u_1(\beta) u_2(\beta) \cdots u_s(\beta)], \qquad U'(\beta)U(\beta) = I.$$

Moreover, by suitable ordering, the mappings $\beta \to \lambda_j(\beta)$ and $\beta \to u_j(\beta)$ can be chosen to be continuous (in fact, analytic) for $\beta \in \mathbb{R}$ and $j \in \underline{s}$ (see, for example, [8, § 2.6]).

From inequality (A3) it follows that for each $\beta > 0$

(A7) $$\begin{aligned} \psi(\beta) &:= \max\{x'M(\beta)x : \|x\| = 1\} \\ &= \max\{\lambda_j(\beta) : j \in \underline{s}\} > 0. \end{aligned}$$

Clearly, ψ is a continuous function of β for all β in $\mathbb{R}$.

Let $0 < \beta_1 < \beta_2 < \cdots < \beta_k < \cdots$ denote the collection of exceptional points to the right of $\beta = 0$ (i.e., points where the graphs of the functions $\lambda_j(.)$ may cross to each other). For a more precise definition see § 2.1.8 of [8]. Note also that the ordering of the exceptional points given above makes sense, since there is always a finite number of them in every compact set of $\mathbb{R}$. It now follows that the function ψ defined in (A7) can be written as

(A8) $$\psi(\beta) = \begin{cases} \lambda_{j_0}(\beta) & \text{if } 0 < \beta \leqq \beta_1, \\ \lambda_{j_k}(\beta) & \text{if } \beta_k < \beta \leqq \beta_{k+1}, k \geqq 1 \end{cases}$$

where $j_k \in \underline{s}$ for all $k \geqq 0$ and $\lambda_{j_0}(\beta)$, $\lambda_{j_k}(\beta)$; $k \geqq 1$, denote the maximum eigenvalues of $M(\beta)$ on the intervals $(0, \beta_1]$ and $(\beta_k, \beta_{k+1}]$, respectively.

From (A7) and (A8) we observe that for each $k \geqq 1$, there exists $\varepsilon_k > 0$ such that both $\lambda_{j_{k-1}}(\beta)$ and $\lambda_{j_k}(\beta)$ are strictly positive on $|\beta - \beta_k| < \varepsilon_k$. Let $\beta_0 < 0$ be given, let $\varepsilon_0 = 0$, and define the following vector-valued function:

(A9) $$x(\beta) := \begin{cases} u_{j_k}(\beta) & \text{if } \beta_k + \varepsilon_k < \beta \leqq \beta_{k+1} - \varepsilon_{k+1}, \\ \dfrac{z_k(\beta)}{\|z_k(\beta)\|} & \text{if } \beta_{k+1} - \varepsilon_{k+1} < \beta \leqq \beta_{k+1} + \varepsilon_{k+1} \end{cases}$$

where $k \geqq 0$ and the functions $z_k(.)$ are defined by

(A10a) $$z_k(\beta) := (1 - g_k(\beta))u_{j_k}(\beta) + g_k(\beta)u_{j_{k+1}}(\beta),$$

(A10b) $$g_k(\beta) := \frac{1}{2\varepsilon_{k+1}}(\beta - (\beta_{k+1} - \varepsilon_{k+1}))$$

where the normal vector $u_{j_k}(\beta)$ is an eigenvector corresponding to $\lambda_{j_k}(\beta)$ (see equation (A6)).

Observing the following, the proof is complete:

(i) The analyticity of $\beta \to u_{j_k}(\beta)$, the continuity of $\beta \to g_k(\beta)$, and the fact that $z_k(\beta) \neq 0$ for all $k \geqq 0$, implies that $\beta \to x(\beta)$, as defined in (A9), is continuous on (β_0, ∞); and

(ii) A trivial computation using (A6)-(A10) shows that $x(\beta)'M(\beta)x(\beta) > 0$, for all $\beta > 0$. Moreover, from (A9) it is clear that $\|x(\beta)\| = 1$ for all $\beta > \beta_0$. □

We are now ready to prove Theorem 3.15.

Proof of Theorem 3.15. Suppose first that $A_3 < 0$, i.e., $\lambda_M(A_3) < 0$. In this case it is straightforward to verify that there always exists $\beta > 0$ (sufficiently small) such that $\beta^2 A_2 + \beta A_1 + A_3 < 0$.

Now, we need to consider the case $\lambda_M(A_3) \geqq 0$. The proof is in two steps. First, we show that there exists $\varepsilon > 0$ such that the symmetric matrices A_1, $\tilde{A}_2 := A_2 + \varepsilon I$, and A_3 satisfy all the assumptions of Lemma A1. Indeed, from assumption (ii) of Theorem 3.15, it follows that

$$\mu := \min \{\delta(x) : \|x\| = 1,\ x'A_3x \geqq 0\} > 0. \tag{A11}$$

(Note that δ is continuous in x and that the intersection of the unit sphere with $x'A_3x \geqq 0$ is a compact set.) Set $\varepsilon > 0$ such that

$$4\lambda_M(A_3)\varepsilon < \mu. \tag{A12}$$

Since μ is strictly positive, such an ε can always be found. Now, consider the function:

$$\tilde{\delta}(x) := (x'A_1x)^2 - 4(x'\tilde{A}_2x)(x'A_3x), \tag{A13}$$

and observe that since $A_2 \geqq 0$ and $\varepsilon > 0$, it follows that $\tilde{A}_2 := A_2 + \varepsilon I$ is positive definite. Hence, from (A13) we conclude that $\tilde{\delta}(x) > 0$ for all $x \neq 0$ such that $x'A_3x < 0$. On the other hand, from (A11) through (A13) we obtain,

$$\begin{aligned}
\tilde{\delta}(x) &= \delta(x) - 4\varepsilon(x'x)(x'A_3x), \\
&\geqq \delta(x) - 4\varepsilon\lambda_M(A_3)(x'x)^2, \\
&\geqq (\mu - 4\varepsilon\lambda_M(A_3))(x'x)^2,
\end{aligned}$$

for all $x \neq 0$ such that $x'A_3x \geqq 0$.

It follows that the matrices A_1, $\tilde{A}_2$, and A_3 satisfy all the assumptions of Lemma A1. Therefore, there exists $\beta > 0$ such that $\beta^2(A_2 + \varepsilon I) + \beta A_1 + A_3 \leqq 0$, which certainly implies $\beta^2 A_2 + \beta A_1 + A_3 \leqq -\varepsilon\beta^2 I < 0$. □

REFERENCES

[1] B. R. BARMISH, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain linear system*, J. Optim. Theory Appl., 46 (1985), pp. 399-408.

[2] J. M. COLLADO AND I. R. PETERSEN, *Correction to "A stabilization algorithm for a class of uncertain linear systems,"* Systems Control Lett., 11 (1988), p. 83.

[3] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State-space solutions to standard H^2 and H^∞ control problems*, in Proc. Automat. Control Conference, Atlanta, GA, 1988.

[4] B. A. FRANCIS AND T. T. GEORGIOU, *Stability theory for linear time-invariant plants with periodic digital controllers*, in Proc. Automat. Control Conference, Atlanta, GA, 1988.

[5] C. V. HOLLOT AND B. R. BARMISH, *Optimal quadratic stabilizability of uncertain linear systems*, in Proc. 18th Allerton Conference on Communication, Control and Computation, University of Illinois, Monticello, IL, 1980.

[6] D. H. JACOBSON, *Extensions of Linear-Quadratic Control, Optimization and Matrix Theory*, Academic Press, London, 1977.

[7] P. T. KABAMBA, *Control of linear systems using generalized sampled-data hold functions*, IEEE Trans. Automat. Control, 32 (1987), pp. 772-783.

[8] T. KATO, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.

[9] P. P. KHARGONEKAR AND K. R. POOLLA, *Uniformly optimal control of linear time-invariant plants: nonlinear time-varying controllers*, Systems Control Lett., 6 (1986), pp. 303-308.

[10] P. P. KHARGONEKAR, T. T. GEORGIOU, AND A. M. PASCOAL, *On the robust stabilizability of linear time-invariant plants with unstructured uncertainty*, IEEE Trans. Automat. Control, 32 (1987), pp. 201-207.

[11] P. P. KHARGONEKAR, I. R. PETERSEN, AND K. ZHOU, *Robust stabilization and H^∞ optimal control*, Internal Report 87-KPZ, Department of Electrical Engineering, University of Minnesota, Minneapolis, MN, 1987. A brief version appeared in the Proc. 26th Allerton Conf. on Communication, Control and Computation, Monticello, IL, 1987.

[12] P. P. KHARGONEKAR, K. R. POOLLA, AND A. TANNENBAUM, *Robust control of linear time-invariant plants by periodic compensation*, IEEE Trans. Automat. Control, 30 (1985), pp. 1088-1096.

[13] S. LEE, S. M. MEERKOV, AND T. RUNOLFSSON, *Vibrational feedback control: zero placement capabilities*, IEEE Trans. Automat. Control, 32 (1987), pp. 604-611.

[14] I. R. PETERSEN, *Quadratic stabilizability of uncertain linear systems: existence of a nonlinear stabilizing control does not imply existence of a linear stabilizing control*, IEEE Trans. Automat. Control, 30 (1985), pp. 291-293.

[15] ———, *A stabilization algorithm for a class of uncertain linear systems*, Systems Control Lett., 8 (1987), pp. 351-357.

[16] ———, *Stabilization of an uncertain linear system in which uncertain parameters enter into the input matrix*, SIAM J. Control Optim., 26 (1988), pp. 1257-1264.

[17] ———, *Quadratic stabilizability of uncertain linear systems containing both constant and time-varying uncertain parameters*, J. Optim. Theory Appl., 57 (1988), pp. 439-461.

[18] I. R. PETERSEN AND B. R. BARMISH, *The stabilization of single input uncertain linear systems via linear control*, in Proc. 6th International Conference on Analysis and Optimization of Systems, Lecture Notes in Control and Information Sciences 62, Springer-Verlag, Berlin, New York, 1984, pp. 69-83.

[19] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear systems*, Automatica, 22 (1986), pp. 397-411.

[20] K. R. POOLLA AND T. TING, *Nonlinear time-varying controllers for robust stabilization*, IEEE Trans. Automat. Control, 32 (1987), pp. 195-200.

[21] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.

[22] E. D. SONTAG, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462-471.

[23] K. ZHOU AND P. P. KHARGONEKAR, *Robust stabilization of linear systems with norm bounded time-varying uncertainty*, Systems Control Lett., 10 (1988), pp. 17-20.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 27, No. 6, pp. 1477-1489, November 1989

014

THE HAMILTON–JACOBI–BELLMAN EQUATION FOR TIME-OPTIMAL CONTROL*

L. C. EVANS† AND M. R. JAMES‡

Abstract. In this paper several assertions concerning viscosity solutions of the Hamilton-Jacobi-Bellman equation for the optimal control problem of steering a system to zero in minimal time are proved. First two rather general uniqueness theorems are established, asserting that any positive viscosity solution of the HJB equation must, in fact, agree with the minimal time function near zero; if also a boundary condition introduced by Bardi [*SIAM J. Control Optim.*, 27 (1988), pp. 776–785] is satisfied, then the agreement is global. Additionally, the Hölder continuity of any subsolution of the HJB equation is proved in the case where the related dynamics satisfy a Hörmander-type hypothesis. This last assertion amounts to a "half-derivative" analogue of a theorem of Crandall and Lions [*Trans. Amer. Math. Soc.*, 277 (1983), pp. 1–42] concerning Lipschitz viscosity solutions.

Key words. Hamilton-Jacobi-Bellman equation, viscosity solution, minimal time function

AMS(MOS) subject classifications. 35F30, 49C20, 49E30

1. Introduction. In this paper we study the *minimum time optimal control problem* for the system

$$\begin{aligned} &\dot{x}(t) = f(x(t), u(t)) \qquad \text{(a.e. } t > 0), \\ &x(0) = x \end{aligned} \tag{1.1}$$

where $x \in \mathbb{R}^n$ and

$$f : \mathbb{R}^n \times U \to \mathbb{R}^n$$

are given, and U denotes a compact subset of, say, $\mathbb{R}^m$. The measurable function

$$u(\cdot) : [0, \infty) \to U$$

is a *control*, and

$$x(\cdot) : [0, \infty) \to \mathbb{R}^n$$

is the corresponding *state*, sometimes written $x(\cdot) = x^u(\cdot)$. We are interested in studying the *minimum time function*

$$T : \mathbb{R}^n \to [0, \infty],$$

defined so that $T(x)$ is the infimum over all controls $u(\cdot)$ of the time taken for the solution of (1.1) to reach the origin.

It is known that under various controllability assumptions T is continuous and thus is a solution of the appropriate *Hamilton-Jacobi-Bellman* (HJB) *equation* in the viscosity sense (see, for example, [6], [9], [11], [15]). In addition, Hermes [6], Sussmann [15], and others have introduced methods for constructing optimal controls via feedback synthesis from appropriate solutions of this HJB equation.

* Received by the editors October 10, 1988; accepted for publication (in revised form) February 3, 1989.

† Department of Mathematics, University of Maryland, College Park, Maryland 20742. The research of this author was supported in part by National Science Foundation grant DMS-86-01532. The author was a part-time member of the Institute for Physical Science and Technology at the University of Maryland, College Park, Maryland 20742, when this paper was written.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by National Science Foundation grant CDR-85-00108.

Herein we prove two uniqueness results. The first asserts that if Ω is a neighborhood of zero and S is a positive viscosity solution of the HJB equation in $\Omega - \{0\}$ satisfying $S(0) = 0$, then there exists another neighborhood Ω' of zero in which S equals the minimal time function T. Quite recently we have seen a new paper of Bardi [2], who has introduced a boundary condition and obtained a uniqueness theorem for the resulting boundary value problem. Our second uniqueness result employs Bardi's boundary condition: If Ω is an open set containing 0 and S is a positive viscosity solution of the HJB equation in $\Omega - \{0\}$ satisfying $S(0) = 0$ and

$$S(x) \to \infty \quad \text{uniformly as } x \to \partial\Omega,$$

then S equals the minimum time function T in Ω. Our proof differs from Bardi's in that we use a representation technique and thereby identify S with T.

These assertions that we prove without any controllability hypotheses, answer a question posed by Hermes in [6]. The point is that if we can find, by any means whatsoever, a positive viscosity solution S of the HJB equation in a region $\Omega - \{0\}$ that vanishes at zero, then S must necessarily be the minimum time function, at least near zero, and consequently the system must in fact be small time locally controllable. Furthermore, if the above boundary condition is satisfied, then Ω coincides with the set C of points controllable to the origin. We see therefore that the explicit solutions constructed for various examples by Hermes in [6] are indeed the minimum time functions.

The proof of these uniqueness assertions appears in § 3, after some preliminaries in § 2.

In § 4 we discuss the regularity of a viscosity subsolution S of the HJB equation in Ω, for the special case that f has the form

$$f(x, u) \equiv \sum_{k=1}^{m} u_k f_k(x) \tag{1.2}$$

where now

$$U = [-1, 1]^m. \tag{1.3}$$

Imposing a simple Hörmander-type requirement concerning the Lie brackets of the vector fields $\{f_j\}_{j=1}^m$, we show that S is locally Hölder continuous with exponent $\frac{1}{2}$. This assertion is certainly clear near zero for viscosity solutions S, in view of § 3 and known regularity theorems for the minimum time function (e.g., Stefani [12], Liverovskii [10]). What we show, in fact, is that the HJB equation itself forces the Hölder continuity.

2. Preliminaries. We now restate our minimum time control problem more precisely as follows.

Let us assume that

$$f: \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$$

is a given smooth function satisfying the Lipschitz condition

$$|f(x, u) - f(y, u)| \leqq K|x - y|$$

for all $x, y \in \mathbb{R}^n$ and all $u \in U$, where $U \subset \mathbb{R}^m$ is compact. Then for any measurable mapping

$$u(\cdot): [0, \infty) \to U,$$

and any point $x \in \mathbb{R}^n$, problem (1.1) has a unique absolutely continuous solution $x(\cdot) = x^u(\cdot)$. We call $u(\cdot)$ an *admissible control* and denote by $\mathcal{U}$ the collection of all admissible controls.

Now let $\Omega \subset \mathbb{R}^n$ be open, with $0 \in \Omega$. We say that the system (1.1) is *controllable to the origin* from $x \in \Omega$ if there exists a control $u(\cdot) \in \mathcal{U}$ and a time

$$0 < \tau_x^u < \infty$$

so that

$$x(\tau_x^u) = 0.$$

Define the following:

$$C(t) \equiv \{x \in \Omega \,|\, \text{there exists a control } u(\cdot) \in \mathcal{U} \text{ with } \tau_x^u \leqq t\};$$

this is the set of states *controllable to the origin within time t.* The set of points controllable to the origin is defined by

$$C \equiv \bigcup_{t>0} C(t).$$

We say that the system (1.1) is *small time locally controllable* (STLC) at the origin if

$$0 \in \operatorname{int} C(t) \quad \text{for each } t > 0.$$

Finally, define the *minimum time function*

$$T(x) \equiv \inf\{\tau_x^u \,|\, u \in \mathcal{U}\}.$$

Then

$$T(0) = 0,$$

and

$$0 < T(x) \leqq +\infty \quad \text{for } x \in \Omega - \{0\}.$$

Furthermore, Sussmann [15] has shown that $T(\cdot)$ is continuous at 0 if and only if (1.1) is STLC at the origin.

The associated dynamic programming *Hamilton-Jacobi-Bellman* equation is the partial differential equation

$$H(x, DS) = 0 \quad \text{in } \Omega - \{0\}$$

where the Hamiltonian

$$H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

is

$$H(x, p) \equiv \max_{u \in U} \{-f(x, u) \cdot p\} - 1 \qquad (x, p \in \mathbb{R}^n). \tag{2.1}$$

3. Uniqueness. Our first uniqueness assertion is Theorem 3.1.

THEOREM 3.1. *Assume* $S \in C(\Omega)$ *satisfies*

$$\begin{aligned} &\text{(a)} \quad S(0) = 0, \\ &\text{(b)} \quad S(x) > 0, \qquad x \in \Omega - \{0\}, \\ &\text{(c)} \quad H(x, DS) = 0 \quad \textit{in } \Omega - \{0\} \textit{ in the viscosity sense.} \end{aligned} \tag{3.1}$$

Then there exists $r > 0$ *such that*

$$S(x) = T(x) \quad \textit{for } x \in B(0, r) \subset \Omega;$$

and, in particular, the system (1.1) *is* STLC *at the origin.*

Note carefully that our sole hypothesis is the existence of a function $S \in C(\Omega)$ verifying (3.1)(a)-(c).

Proof. (1) We start by fixing some bounded smooth open set $\Omega' \subset\subset \Omega$ with $0 \in \Omega'$. Defining then

$$g(x) \equiv S(x) \qquad (x \in \partial\Omega'),$$

we see that $S \in C(\bar{\Omega}')$ is a viscosity solution of

$$H(x, DS) = 0 \quad \text{in } \Omega' - \{0\}, \qquad S = g \quad \text{on } \partial\Omega'. \tag{3.2}$$

Now set

$$\Phi(r) \equiv -\log(1-r) \qquad (r < 1), \tag{3.3}$$

$$\Psi(s) \equiv \Phi^{-1}(s) = 1 - e^{-s} \qquad (-\infty < s < \infty). \tag{3.4}$$

Write

$$R(x) \equiv \Psi(S(x)) \qquad (x \in \Omega'), \tag{3.5}$$

$$h(x) \equiv \Psi(g(x)) \qquad (x \in \partial\Omega'). \tag{3.6}$$

We note that clearly $R \in C(\bar{\Omega}')$ satisfies

$$R(0) = 0, \tag{3.7}$$

$$R(x) = h(x) \qquad (x \in \partial\Omega'). \tag{3.8}$$

In addition, we claim that

$$R + H(x, DR) = 0 \quad \text{in } \Omega' - \{0\} \quad \text{in the viscosity sense.} \tag{3.9}$$

To see this we note that Crandall, Evans, and Lions [4, Prop. 1.2] imply that $R = \Psi(S)$ solves

$$\tilde{H}(x, R, DR) = 0 \quad \text{in } \Omega' - \{0\} \tag{3.10}$$

in the viscosity sense, for

$$\tilde{H}(x, t, p) \equiv H\left(x, \frac{p}{1-r}\right).$$

But since $0 \leqq R < 1$, we have further that

$$(1-R)\tilde{H}(x, R, DR) = R + H(x, DR) = 0 \quad \text{in } \Omega' - \{0\}$$

in the viscosity sense, as required.

Note also now that R is the unique solution of (3.7)-(3.9), in view of Theorem III.1 in Crandall and Lions [3].

(2) We now obtain a control theoretic representation formula for R (and thus for S) in Ω'. For each point $x \in \Omega' - \{0\}$ and control $u(\cdot) \in \mathcal{U}$ define

$$\sigma_x^u \equiv \inf\{t > 0 \,|\, x(t) \in \{0\} \cup \partial\Omega'\}$$

where as usual $x(\cdot) = x^u(\cdot)$ is the solution of system (1.1) corresponding to the control $u(\cdot)$.

Now write the value function

$$V(x) \equiv \inf_{u \in \mathcal{U}} \left\{ \int_0^{\sigma_x^u} e^{-t}\, dt + e^{-\sigma_x^u} \chi_{\{x(\sigma_x^u) \in \partial\Omega'\}} h(x(\sigma_x^u)) \right\}. \tag{3.11}$$

We assert that then

$$R(x) = V(x) \quad \text{in } \bar{\Omega}'. \tag{3.12}$$

This equality follows from almost exactly the same proofs as in Lions [9] or Evans and Ishii [5, Thm. 4.1].

(3) We now claim that there exists $r>0$ such that

$$R(x)=\inf_{u\in\mathcal{U}}\{\Psi(\tau_x^u)\} \qquad (x\in B(0,r)\subset\Omega'). \tag{3.13}$$

To verify this we set

$$s\equiv\inf\{|y|\,|\,y\in\partial\Omega'\},$$
$$\|f\|\equiv\sup\{|f(x,u)|\,|\,x\in\Omega',\ u\in U\},$$
$$\|R\|_r\equiv\sup\{|R(x)|\,|\,|x|\leqq r\},$$
$$\tau_{x,y}^u\equiv\inf\{t>0\,|\,x^u(0)=x,\ x^u(t)=y\}\leqq\infty.$$

We denote by $\mathcal{U}_{x,y}$ the set of admissible controls with $\tau_{x,y}^u<\infty$.

Finally, let us write

$$U(x,y)\equiv\inf\{\tau_{x,y}^u\,|\,u(\cdot)\in\mathcal{U}_{x,y}\}.$$

Choose $0<r<s$ so small that

$$B(0,r)\subset\Omega'$$

and

$$\|R\|_r\leqq\frac{1}{2}\Psi\left(\frac{s-r}{2\|f\|}\right). \tag{3.14}$$

We must verify (3.13). Let us choose $x\in B(0,r)$, $y\in\mathbb{R}^n-B(0,s)$, and suppose $U(x,y)<\infty$. Then we select a control $u(\cdot)\in\mathcal{U}_{x,y}$ satisfying

$$U(x,y)\leqq\tau_{x,y}^u\leqq U(x,y)+\frac{s-r}{2\|f\|}. \tag{3.15}$$

In view of (1.1) we have

$$y-x=\int_0^{\tau_{x,y}^u}f(x(s),u(s))\,ds;$$

whence

$$s-r\leqq|y-x|\leqq\tau_{x,y}^u\|f\|. \tag{3.16}$$

Consequently, (3.15) implies

$$U(x,y)\geqq\frac{s-r}{2\|f\|} \quad \text{if } |x|\leqq r, \quad |y|\geqq s. \tag{3.17}$$

Now fix

$$0<\delta<\frac{1}{2}\Psi\left(\frac{s-r}{2\|f\|}\right). \tag{3.18}$$

Then if $x\in B(0,r)$, there exists a control $u(\cdot)\in\mathcal{U}$ with

$$\begin{aligned}\int_0^{\sigma_x^u}e^{-t}\,dt+e^{-\sigma_x^u}\chi_{\{x(\sigma_x^u)\in\partial\Omega'\}}h(x(\sigma_x^u))&\leqq V(x)+\delta\\ &=R(x)+\delta \quad \text{by (3.12).}\end{aligned} \tag{3.19}$$

First assume that

$$y \equiv x(\sigma_x^u) \in \partial\Omega'. \tag{3.20}$$

Then since

$$\Psi(s) = 1 - e^{-s} = \int_0^s e^{-t}\, dt,$$

it follows that

$$\Psi(\sigma_x^u) + e^{-\sigma_x^u} h(x(\sigma_x^u)) \leqq R(x) + \delta < \Psi\left(\frac{s-r}{2\|f\|}\right),$$

according to (3.14), (3.18). Since $h > 0$ and Ψ is strictly increasing, it follows that

$$\begin{aligned} \sigma_x^u &< \frac{s-r}{2\|f\|} \\ &\leqq U(x, y) \quad \text{by (3.17)}, \end{aligned}$$

a contradiction. Consequently, we must have

$$y \equiv x(\sigma_x^u) = 0$$

in (3.19). Hence (3.11), (3.12), and (3.19) imply that if $x \in B(0, r)$, then

$$\begin{aligned} R(x) = V(x) &= \inf_{u \in \mathcal{U}} \left\{ \int_0^{\sigma_x^u} e^{-t}\, dt \,|\, x^u(\sigma_x^u) = 0 \right\} \\ &= \inf_{u \in \mathcal{U}} \{\Psi(\tau_x^u)\}. \end{aligned}$$

This is formula (3.13).

(4) Finally, due to (3.5) and (3.13), we see

$$\begin{aligned} S(x) &= \Phi(R(x)) \\ &= \inf_{u \in \mathcal{U}} \{\tau_x^u\} \\ &= T(x) \end{aligned}$$

if $x \in B(0, r)$. □

Now we turn to our second uniqueness result that incorporates the boundary condition introduced by Bardi [2]. For $S \in C(\Omega)$ we say that

$$S(x) \to \infty \quad \text{uniformly as } x \to \partial\Omega$$

provided that for all $M > 0$ there exists $\delta > 0$ such that

$$S(x) \geqq M$$

provided $|x| \geqq 1/\delta$ or dist. $(x, \partial\Omega) \leqq \delta$.

THEOREM 3.2. *Assume $S \in C(\Omega)$ satisfies*

$$\begin{aligned} &\text{(a)} \quad S(0) = 0, \\ &\text{(b)} \quad S(x) > 0 \qquad (x \in \Omega - \{0\}), \\ &\text{(c)} \quad S(x) \to \infty \quad \textit{uniformly as } x \to \partial\Omega, \\ &\text{(d)} \quad H(x, DS) = 0 \quad \textit{in } \Omega - \{0\} \textit{ in the viscosity sense.} \end{aligned} \tag{3.21}$$

Then

$$S(x) = T(x) \quad \textit{for all } x \in \Omega,$$

and hence $\Omega = C$.

The proof is similar to that of Theorem 3.1, and so we indicate only the essential changes.

Proof. Defining

$$R(x) = \Psi(S(x)) \qquad (x \in \Omega)$$

we see that $R \in C(\Omega)$ satisfies

$$R + H(x, DR) = 0 \quad \text{in } \Omega - \{0\} \text{ in the viscosity sense,}$$

and also

$$R(0) = 0, \qquad 0 < R(x) < 1 \quad (x \in \partial\Omega - \{0\}). \tag{3.22}$$

In view of the boundary condition (3.21)(c), R can be uniquely extended to a function $R \in BUC(\bar{\Omega})$ by setting (cf. Bardi [2])

$$R(x) = 1 \qquad (x \in \partial\Omega).$$

Employing the methods of Lions [9] or Evans and Ishii [5], we obtain the following control theoretic representation. Define

$$\sigma_x^u \equiv \inf\{t > 0 \,|\, x(t) \in \{0\} \cup \partial\Omega\};$$

then

$$R(x) = \inf_{u \in \mathcal{U}} \left\{ \int_0^{\sigma_x^u} e^{-t}\, dt + e^{-\sigma_x^u} \chi_{\{x(\sigma_x^u) \in \partial\Omega\}} \right\}$$

for all $x \in \bar{\Omega}$.

In fact, we claim that

$$R(x) = \inf_{u \in \mathcal{U}} \{\Psi(\tau_x^u)\} \quad \text{for all } x \in \Omega. \tag{3.23}$$

To see this, note that if $x \in \Omega$, then (3.22) implies

$$\alpha \equiv 1 - R(x) > 0.$$

Choose $u \in \mathcal{U}$ such that

$$\int_0^{\sigma_x^u} e^{-t}\, dt + e^{-\sigma_x^u} \chi_{\{x(\sigma_x^u) \in \partial\Omega\}} \leqq R(x) + \frac{\alpha}{2}. \tag{3.24}$$

Suppose now that $x(\sigma_x^u) \in \partial\Omega$. Then (3.24) implies

$$1 = \int_0^{\sigma_x^u} e^{-t}\, dt + e^{-\sigma_x^u} \leqq R(x) + \frac{\alpha}{2},$$

a contradiction. Consequently, $x(\sigma_x^u) = 0$, and (3.23) follows. We conclude by noting that

$$S(x) = \Phi(R(x)) = T(x) \qquad (x \in \Omega). \qquad \square$$

4. Regularity.

Motivation. Let us now address the problem of the smoothness of solutions to the HJB equation in the case that f has the explicit form

$$f(x, u) = \sum_{k=1}^{m} u_k f_k(x), \tag{4.1}$$

where

$$f_k : \mathbb{R}^n \to \mathbb{R}^n \qquad (k = 1, \cdots, m)$$

are given smooth functions, and

$$u = (u_1, \cdots, u_m) \in U \equiv [-1, 1]^m.$$

In this case the corresponding HJB equation reads

$$\sum_{k=1}^{m} |f_k(x) \cdot DS| = 1 \quad \text{in } \Omega - \{0\}. \tag{4.2}$$

For heuristic purposes, let us for the moment suppose that S is a smooth solution of (4.2). Then (4.2) implies

$$|f_k \cdot DS| \leqq 1 \qquad (k = 1, \cdots, m) \tag{4.3}$$

in $\Omega - \{0\}$, so that the rate of change of S in the direction f_k is bounded ($k = 1, \cdots, m$). Thus, if

$$\text{span}\{f_k(x) \mid k = 1, \cdots, m\} = \mathbb{R}^n \tag{4.4}$$

for some point $x \in \Omega - \{0\}$, then we can derive from (4.2) an estimate on $|DS(x)|$. On the other hand, suppose (4.4) fails. Then we cannot generally hope to estimate $|DS(x)|$: we can deduce from (4.2) estimates only on the components of $DS(x)$ in the direction $f_1(x), \cdots, f_m(x)$.

Let us now however assume the *Hörmander condition* instead of (4.4):

$$\text{span}\{f_i(x), [f_j, f_k](x) \mid i, j, k = 1, \cdots, m\} = \mathbb{R}^n \tag{4.5}$$

for each $x \in \Omega - \{0\}$, where the pairing "[,]" denotes the usual Lie bracket. It is then well known from control theory that the minimum time function $T(x)$ is Hölder continuous with exponent $\frac{1}{2}$. In view of § 3 above it therefore seems reasonable to expect that our solution S of (4.2) will have this same regularity.

Below we show, more generally, that if (4.5) holds and S satisfies, say, (4.3) in the viscosity sense, then S is locally Hölder continuous with exponent $\frac{1}{2}$. This is a kind of "half-derivative" analogue of a theorem of Crandall and Lions [3] to the effect that if S is a viscosity solution of the partial differential equation

$$H(x, DS) = 0 \quad \text{in } \Omega \tag{4.6}$$

where

$$\lim_{|p| \to \infty} H(x, p) = +\infty,$$

then S is Lipschitz.

THEOREM 4.1. *Let C be a positive constant. Assume $S \in C(\Omega)$ satisfies*

$$|f_k \cdot DS| \leqq C \quad \text{in } \Omega \tag{4.7}$$

in the viscosity sense $(k=1,\cdots,m)$, *where the smooth vector fields* $\{f_k\}_{k=1}^m$ *satisfy condition* (4.5) *for each point* $x\in\Omega$. *Then*

$$S\in C^{0,1/2}_{\text{loc}}(\Omega).$$

Remark. More precisely, our hypothesis is that for each smooth function ϕ,

(4.8) If $S-\phi$ has a maximum at a point $x_0\in\Omega$, then $|f_k(x_0)\cdot D\phi(x_0)|\leqq C$ for $k=1,\cdots,m$.

Proof. (1) We first regularize our function S by setting for each $\varepsilon>0$

$$S^\varepsilon(x)=\sup_{y\in\Omega}\left[S(y)-\frac{1}{\varepsilon}|x-y|^2\right]\qquad(x\in\Omega). \tag{4.9}$$

This is the Yosida–Moreau sup-convolution, the importance of which for viscosity solution is discussed by Jensen, Lions, and Souganidis [7] (cf. also Lasry and Lions [8]). We easily check that

$$S^\varepsilon\in W^{1,\infty}_{\text{loc}}(\Omega).$$

Additionally, for each open set $\Omega'\subset\subset\Omega$, we have the estimate

$$|DS^\varepsilon|\leqq\frac{C(\Omega')}{\varepsilon^{1/2}}\quad\text{a.e. in }\Omega', \tag{4.10}$$

the constant $C(\Omega')$ depending only on S and dist $(\Omega',\partial\Omega)$.

(2) We claim now that for each sufficiently small $\varepsilon>0$ we have the estimate

$$|f_k(x)\cdot DS^\varepsilon(x)|\leqq C(\Omega')\quad\text{a.e. in }\Omega', \tag{4.11}$$

for some constant $C(\Omega')$ depending only on S and dist. $(\Omega',\partial\Omega)$. To see this, choose $\Omega'\subset\subset\Omega''\subset\subset\Omega$, and suppose that ϕ is a smooth function and that

$$S^\varepsilon-\phi\text{ has a local maximum at some point }x_0\in\Omega'. \tag{4.12}$$

We then deduce as in [7] that

$$S-\tilde\phi\text{ has a local maximum at }y_0,$$

for

$$\tilde\phi(y)\equiv\phi(y-y_0+x_0),$$

with y_0 selected so that

$$S^\varepsilon(x_0)=S(y_0)-\frac{1}{\varepsilon}|x_0-y_0|^2. \tag{4.13}$$

In view of (4.8) then

$$|f_k(y_0)\cdot D\tilde\phi(y_0)|=|f_k(y_0)\cdot D\phi(x_0)|\leqq C.$$

But now (4.12) implies

$$|x_0-y_0|\leqq C\varepsilon^{1/2},$$

C depending only on S and Ω''. Hence for each $k=1,\cdots,m$

$$\begin{aligned}|f_k(x_0)\cdot D\phi(x_0)|&\leqq C+|f_k(y_0)-f_k(x_0)||D\phi(x_0)|\\&\leqq C+C\varepsilon^{1/2}|D\phi(x_0)|\\&\leqq C,\end{aligned} \tag{4.14}$$

since (4.10) and (4.12) imply

$$|D\phi(x_0)| \leqq \frac{C}{\varepsilon^{1/2}}.$$

The constant C in (4.14) depends only on S and dist. $(\Omega', \partial\Omega)$, and so (4.11) is proved.

(3) Now fix any index $k \in \{1, \cdots, m\}$ and any point $x \in \Omega$, and consider the ordinary differential equation

$$\dot{x}(s) = f_k(x(s)), \quad x(0) = x \quad (s \in \mathbb{R}, x \in \Omega). \tag{4.15}$$

For later notational convenience, let us write

$$x(s) \equiv X^k(s)x \qquad (s \in \mathbb{R}, x \in \Omega) \tag{4.16}$$

to display the dependence on the vector field f_k and the initial point x.

(4) We now claim that if $z \in \Omega'$, then

$$|S^\varepsilon(z) - S^\varepsilon(X^k(t)z)| \leqq C|t| \qquad (k = 1, \cdots, m) \tag{4.17}$$

for some constant C and all sufficiently small $|t|$, $\varepsilon > 0$. To see this, note that we may as well assume $f_k(z) \neq 0$, since otherwise (4.17) is obvious. Now since S^ε is Lipschitz, Rademacher's Theorem implies that $DS^\varepsilon(y)$ exists for almost every $y \in \Omega$. Consequently by the Coarea Formula, we can find a sequence of points

$$z_l \to z,$$

and a sufficiently small time $|t| > 0$ such that

$$DS^\varepsilon \text{ exists and (4.11) holds at } x = X^k(s)z_l \text{ for a.e. } -|t| < s < |t|, \tag{4.18}$$

the a.e. (almost every) taken with respect to one-dimensional Lebesgue measure. Now for each l the mapping

$$s \mapsto S^\varepsilon(X^k(s)z_l)$$

is Lipschitz, with

$$\frac{d}{ds} S^\varepsilon(X^k(s)z_l) = DS^\varepsilon(X^k(s)z_l) \cdot f_k(X^k(s)z_l)$$

for a.e. $-|t| < s < |t|$, according to (4.15) and (4.18). Thus (4.18) and (4.11) imply

$$|S^\varepsilon(z_l) - S^\varepsilon(X^k(t)z_l)| \leqq C|t|.$$

Let l approach infinity to obtain (4.17).

(5) Now fix any point $x_0 \in \Omega'$. Because of the Hörmander condition (4.5) we can select $0 \leqq d \leqq 2m$ and indices $k_j \in \{1, \cdots, m\}$ $(j = 1, \cdots, d)$ such that the vectors

$$f_{k_1}(x_0), \cdots, f_{k_l}(x_0), \qquad [f_{k_{l+1}}, f_{k_{l+2}}](x_0), \cdots, [f_{k_{d-1}}, f_{k_d}](x_0)$$

are a basis of $\mathbb{R}^n$. On relabeling, if necessary, we may as well assume that

$$f_1(x_0), \cdots, f_l(x_0), \qquad [f_{a(l+1)}, f_{b(l+1)}](x_0), \cdots, [f_{a(n)}, f_{b(n)}](x_0)$$

form a basis, where $1 \leqq l \leqq n$ and

$$a, b : \{l+1, \cdots, n\} \to \{1, \cdots, m\} \tag{4.19}$$

are appropriate functions.

(4.20) By continuity the vectors

$$f_1(y), \cdots, f_l(y), \qquad [f_{a(l+1)}, f_{b(l+1)}](y), \cdots, [f_{a(n)}, f_{b(n)}](y)$$

also form a basis of $\mathbb{R}^n$, for each point y sufficiently close to x_0.

(6) Following Strichartz [13], let us now extend the notation introduced in (4.15), (4.16) by writing

$$X^{k,l}(s)x \equiv \begin{cases} X^l(-s^{1/2})X^k(-s^{1/2})X^l(s^{1/2})X^k(s^{1/2})x & \text{if } s>0, \\ x \quad \text{if } s=0, \\ X^k(-|s|^{1/2})X^l(-|s|^{1/2})X^k(|s|^{1/2})X^l(|s|^{1/2})x & \text{if } s<0. \end{cases} \tag{4.21}$$

A lengthy but well-known calculation reveals that

$$\frac{d}{ds} X^{k,l}(s)x\big|_{s=0} = [f_k, f_l](x). \tag{4.22}$$

Given now $t=(t_1, \cdots, t_n) \in \mathbb{R}^n$, set

$$y = \Phi(t) = X^{(a)n,b(n)}(t_n) \cdots X^{a(l+1),b(l+1)}(t_{l+1})X^l(t_l) \cdots X^2(t_2)X^1(t_1)x_0, \tag{4.23}$$

the mappings $a(\cdot)$, $b(\cdot)$ as in (4.19). Then

$$\Phi : \mathbb{R}^n \to \mathbb{R}^n$$

is C^1, and

$$\Phi(0) = x_0.$$

Additionally, using (4.20) and (4.22), we compute

$$\det D\Phi(0) \neq 0.$$

Invoking the Inverse Function Theorem, we deduce that Φ is a C^1 diffeomorphism of some neighborhood U of $t=0$ onto the open ball $U(x_0, r_0)$ for some $r_0>0$. Furthermore, there exists a constant C such that

$$|t| \leqq C|\Phi(t) - x_0| \qquad (t \in U).$$

(7) Now choose any $y \in U(x_0, r_0)$ and set

$$r = |y - x_0|. \tag{4.24}$$

By the above there exists a unique point $t \in U$ with

$$y = \Phi(t)$$

and

$$|t| \leqq Cr. \tag{4.25}$$

We inductively set

$$\begin{aligned} y_0 &= x_0 \\ y_1 &= X^1(t_1)y_0 \\ &\vdots \\ y_l &= X^l(t_l)y_{l-1}, \end{aligned} \tag{4.26}$$

$$
\begin{aligned}
y_{l+1} &= X^{a(l+1),b(l+1)}(t_{l+1})y_l \\
&\vdots \\
y_n &= X^{a(n),b(n)}(t_n)y_{n-1} = y.
\end{aligned} \tag{4.27}
$$

Then for $j = 1, \cdots, l$,

$$
\begin{aligned}
|S^\varepsilon(y_j) - S^\varepsilon(y_{j-1})| &\leqq C|t_j| \quad \text{by (4.17)} \\
&\leqq C|t|.
\end{aligned} \tag{4.28}
$$

Now assume $l+1 \leqq j \leqq n$. Then assuming for simplicity that $t_j > 0$, let us write

$$
\begin{aligned}
y_j^0 &= X^{a(t_j)}(-t_j^{1/2})X^{b(t_j)}(-t_j^{1/2})X^{a(t_j)}(t_j^{1/2})X^{b(t_j)}(t_j^{1/2})y_{j-1} = y_j, \\
y_j^1 &\equiv X^{b(t_j)}(-t_j^{1/2})X^{a(t_j)}(t_j^{1/2})X^{b(t_j)}(t_j^{1/2})y_{j-1}, \\
y_j^2 &\equiv X^{a(t_j)}(t_j^{1/2})X^{b(t_j)}(t_j^{1/2})y_{j-1}, \\
y_j^3 &\equiv X^{b(t_j)}(t_j^{1/2})y_{j-1}, \qquad y_j^4 \equiv y_{j-1}.
\end{aligned}
$$

In view of (4.17) again

$$
\begin{aligned}
|S^\varepsilon(y_j^s) - S^\varepsilon(y_{j-1}^{s+1})| &\leqq C|t_j|^{1/2} \\
&\leqq C|t|^{1/2} \quad (s = 0, 1, 2, 3).
\end{aligned}
$$

Thus

$$
|S^\varepsilon(y_j) - S^\varepsilon(y_{j-1})| \leqq C|t|^{1/2} \tag{4.29}
$$

for $j = l+1, \cdots, n$. Combining at last (4.28) and (4.29) we find

$$
\begin{aligned}
|S^\varepsilon(y) - S^\varepsilon(x_0)| &\leqq C(|t| + |t|^{1/2}) \\
&\leqq C|y - x_0|^{1/2} \quad \text{by (4.24), (4.25).}
\end{aligned} \tag{4.30}
$$

(8) Consequently S^ε is Hölder continuous with exponent $\frac{1}{2}$ at the point x_0. Furthermore, an analysis of the proof shows that we can select the constant C in (4.30) to hold also if x_0 is replaced by any nearby point x. Consequently, S^ε is Hölder continuous with exponent $\frac{1}{2}$ in some neighborhood of x_0, and so by compactness $S^\varepsilon \in C^{0,1/2}_{\text{loc}}(\Omega)$. Since the behavior of S^ε enters the proof only via inequality (4.17), where the constant C is independent of ε, we have

$$
\sup_{\varepsilon > 0} \|S^\varepsilon\|_{C^{0,1/2}(\Omega')} < \infty
$$

for each $\Omega' \subset\subset \Omega$. Since

$$
S^\varepsilon \to S
$$

as $\varepsilon \searrow 0$, we deduce that

$$
S \in C^{0,1/2}_{\text{loc}}(\Omega),
$$

as required. □

Remark. Our proof presumably extends to show that S is locally Hölder continuous with exponent k^{-1}, provided we suppose instead of (4.5) that the k-fold interated Lie brackets of the vector fields $\{f_j(x)\}_{j=1}^m$ span $\mathbb{R}^n$ at each point $x \in \Omega$ $(k = 3, 4, \cdots)$. We have not, however, attempted to work out the relevant details. □

REFERENCES

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

[2] M. BARDI, *A boundary value problem for the minimum time function*, SIAM J. Control Optim., 27 (1989), pp. 776-785.

[3] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[4] 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.

[5] L. C. EVANS AND H. ISHII, *Differential games and nonlinear first order* PDE *on bounded domains*, Manuscripta Math., 49 (1984), pp. 109-137.

[6] H. HERMES, *Feedback synthesis and positive, local solutions to Hamilton-Jacobi-Bellman equations*, in Analysis and Control of Nonlinear Systems, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988, pp. 155-164.

[7] R. JENSEN, P. L. LIONS, AND P. E. SOUGANIDIS, *A uniqueness result for viscosity solutions of second-order fully nonlinear partial differential equations*, to appear.

[8] J. M. LASRY AND P. L. LIONS, *A remark on regularization on Hilbert spaces*, Israel J. Math., 55 (1986), pp. 257-266.

[9] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, New York, 1982.

[10] A. A. LIVEROVSKII, *Hölder property of Bellman Functions of plane control systems*, Differentsial'nye Uravneniya, 17 (1981), pp. 604-613.

[11] V. STAICU, *Minimal time function and viscosity solutions*, preprint.

[12] G. STEFANI, *Polynomial approximations to control systems and local controllability*, in Proc. IEEE 24th Conference on Decision and Control, Ft. Lauderdale, FL, December 1985, pp. 33-38.

[13] R. S. STRICHARTZ, *Sub-Riemannian geometry*, J. Differential Geometry, 24 (1986), pp. 221-263.

[14] H. J. SUSSMANN, *The structure of time optimal trajectories for single-input systems in the plane: the C^∞ nonsingular case*, SIAM J. Control Optim., 25 (1987), pp. 432-465.

[15] ———, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158-194.